SPATIAL-TEMPORAL VARIATIONS AND RISK DETERMINANTS OF VIBRIO PARAHAEMOLYTICUS INFECTIONS IN WASHINGTON STATE

by

ZHE SUN

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland

December 2017

ABSTRACT

Vibrio parahaemolyticus is a naturally-occurring halophilic and asporogenous gram-negative bacterium widely distributed in marine environments and frequently detected in shellfish, particularly oysters. *V. parahaemolyticus* can secrete pathogenic proteins to cause self-restricted gastroenteritis combined with several other related illness symptoms, including but not limited to diarrhea, nausea, vomiting, and even septicemia. *V. parahaemolyticus* infections, also known as *vibriosis*, primarily occur due to consumption of undercooked contaminated shellfish, and have become an unneglectable health concern especially in regions where seafood harvesting activities are frequently undertaken, such

as the estuarine and coastal waters of Washington State. Since it is impossible to eliminate *V. parahaemolyticus* from estuarine environments, necessary research of *vibriosis* such as space-time high-risk cluster detection and screening of risk determinants will help alleviate threats to the public's health.

The ongoing *V. parahaemolyticus* monitoring project in Washington State was launched and has been conducted by the Washington Department of Health since 2008, mainly including microbial genetic parameter determination, environmental and oceanographic risk factor monitoring, and single-source *vibriosis* case trace-back confirmed by epidemiological interviews. Only data collected during 2013-2016 were included into this study in consideration of the deficiencies of epidemiological documenting in the beginning years of the project. A geographical nearest-neighbor re-sampling approach was used for risk parameter interpolation for cases, after which binary logistic regressions were applied to the cases and controls for identifying statistically significant risk factors of *vibriosis*. Spatial statistics and longitudinal analyses were used

to assess the spatial-temporal dependences and variations of *vibriosis* odds.

Vibriosis odds did not increase significantly between 2013 and 2016, while cases were clustered specifically in several regions including *Totten Inlet*, *Bay Center*, and *Nahcotta*. Three environmental factors: *ambient air temperature* (OR=1.07), *surface water temperature* (OR=1.62), and *water salinity* (OR=1.62) contributed positively to the *vibriosis* odds with a negative interaction term observed between water temperature and salinity (OR=0.98). Three genetic factors: *tlh* (OR=0.46), *tdh* (OR=3.70), and *trh* (OR=1.51) most probable numbers were all responsible for odds, with a strong negative interaction term between *tlh* and *tdh* (OR=0.59). This effect modifier was important in exploring the association between *vibriosis* odds and *tlh-tdh* ratios. Although *vibriosis* odds varied between Puget Sound Estuary and Pacific Coastal Areas, risk associations were consistent across years and regions. In summary, this study is one of the very few research projects on *vibriosis* odds, environmental and microbial risk determinants, as well as spatial and temporal variations, which will be helpful in future infection prevention.

Keywords: *Vibrio parahaemolyticus, spatial statistics, temporal analysis, nearest neighbor resampling, risk determinants.*

THESIS ADVISORS:

Dr. Frank C. Curriero, Dr. Benjamin J. Davis, Dr. Ernst W. Spannhake.

CONTENTS

TABLE DIRECTORY

FIGURE DIRECTORY

I INTRODUCTION

Vibrio parahaemolyticus is a type of naturally-occurring halophilic and asporogenous gram-negative bacterium, which is widely distributed in seawater environments and frequently detected in marine organisms like shellfish, particularly the oyster (Davis et al. 2017; Su and Liu 2007; Fujino, Sakazaki, and Tamura 1974). The biological structure of *V. parahaemolyticus* is different from other vibrios like *V. cholerae*, which are of curved rods with a singular flagellum, while *V. parahaemolyticus* is of a straight rod with a number of cilia and a flagellum (Cai, Han, and Wang 2006; Su and Liu 2007). It is therefore theoretically feasible to distinguish *V. parahaemolyticus* from other *vibrios* through

microscopes given their structural divergences. However, a more strict and convincing approach has been developed to isolate and identify *V. parahaemolyticus* by colony cultivation and testing on appropriate culture media such as thiosulfate citrate bile salts sucrose (TCBS) agar (Hara-Kudo et al. 2001; Letchumanan, Chan, and Lee 2014). A two-step enrichment process instead of the traditional one-step salt polymyxin broth based treatments is recommended before plating the bacterium colony on culture media, which improves sensitivity and accuracy of the method (Hara-Kudo et al. 2001).

V. parahaemolyticus is often associated with seafood-borne diseases, especially the serotype of O3:K6 which is of higher infective risk than other serotypes confirmed by previous relative studies, although not every strain of the bacterium can cause infections (Davis et al. 2017; Bag et al. 1999; Nishibuchi and Kaper 1995). Some genetic markers may indicate virulency of the bacterium, such as the ability to secrete virulent proteins including thermostable direct hemolysin (TDH) and TDH-related hemolysin (TRH) (Su and Liu 2007; Miyamoto et al. 1969; Kaneko and Colwell 1973; Honda et al. 1987; Honda,

Ni, and Miwatani 1988).

TDH can cause human gastroenteritis by creating cellular structural modifications on Caco-2 cell monolayers, the key structures for ion exchange in intestinal epithelial cells, effects of which vary from enterotoxicity to cytotoxicity according to the concentrations of TDH proteins (Raimondi et al. 2000). Mechanically, TDH can form porins (Honda et al. 1992) on the plasma membrane of enterocytes and thus lead to the unexpected and non-selective influx of ions (Raimondi et al. 2000). When the concentrations of TDH are low and correspondingly the TDH-generated channels are limited, the unexpected influxes of different species of ions are still within the cells' homeostatic capacity which can help counterbalance the ion exchange flows and maintain the intracellular calcium concentrations without affecting cell functions and viability (Fabbri et al. 1999; Raimondi et al. 1995). On the contrary, at high TDH concentrations, the number of TDH-transformed channels could be elevated to an overwhelming level that exceeds the cell's self-adjusting capacity and thus the negative consequences are uncontrollable massive nonspecific ionic

influx, which can eventually cause irreversible osmotic swelling, cell rounding, or even death (Raimondi et al. 2000). Under that condition, enterocytes could be functionally destroyed and hence lose the defense ability of the epithelial barrier and consequently allow the vibrios to invade the host (Raimondi et al. 2000). What's worse, cytotoxic mechanisms can create a positive feedback that can develop into a blind loop syndrome through impaired luminal clearing and increasing bile acids concentrations, which synergizes the enhancement of TDH production by providing a suitable surrounding (Raimondi et al. 2000). TRH is related with TDH secretion, and is likely to be associated with gastroenteritis in a similar manner.

Besides TDH and TRH, thermolabile hemolysin (TLH) has been verified as a reliable biomarker for *V. parahaemolyticus* species since the protein can be detected in all strains regardless of whether they are virulent or not (Su and Liu 2007). Therefore, genetic markers including *tlh, tdh and trh* have been used for *V. parahaemolyticus* abundance estimation and pathogenicity assessment, which can be detected through oligonucleotide probes

(Kaper et al. 1984; Nishibuchi et al. 1986).

Clinical and epidemiological studies on *V. parahaemolyticus* infections, also known as *vibriosis*, and related food-borne disease outbreaks show that the bacterium is of high morbidity but of quite low mortality. Since *V. parahaemolyticus* cannot be eliminated from marine and estuarine environments, relevant research is needed to mitigate threats to the public's health. The first case identified to be caused by *V. parahaemolyticus* was in Osaka, Japan in 1951 (McCarthy et al. 1999; Daniels, Ray, et al. 2000) which caused 272 illnesses and 20 deaths due to consumption of raw sardines. Self-restricted gastroenteritis is the most common symptom of *vibriosis*, but life-threatening septicemia is also possible, though the probability of occurrence is rather low (Daniels, MacKinnon, et al. 2000). Acute gastroenteritis symptoms include but are not limited to diarrhea, vomiting, headache, abdominal cramps, and nausea (Dadisman et al. 1973; Blake, Weaver, and Hollis 1980; Lozano-Leon et al. 2003; Su and Liu 2007). Fatal cases are often attributed to underlying health conditions of alcoholism or liver disease since it has been observed that these

vulnerable populations are more likely to develop primary septicemia (Daniels, MacKinnon, et al. 2000). It was estimated that around 8% of the *V. parahaemolyticus* infections could result in septicemia based on a study in Florida (Hlady and Klontz 1996).

V. parahaemolyticus is reported to be a leading cause of seafood-borne diseases through consumption of raw or undercooked oysters in the United States (Davis et al. 2017; Bag et al. 1999; Molenda et al. 1972; Daniels, MacKinnon, et al. 2000), as well as in Asian (Chen, Liu, and Zhang 1991; Deepanjali et al. 2005; Liu et al. 2004) and other countries and regions globally. Generally speaking, for the United States, outbreaks occur sporadically across states including Washington, New York, Oregon, Connecticut, California, New Jersey, Texas and others, mostly during summer months (CDC 1998; DePaola et al. 2000; CDC 1999). For example, three outbreaks of 425 gastroenteritis cases due to crab consumption occurred in Maryland in 1971 (Molenda et al. 1972), and 40 outbreaks took place from 1973 to 1998 in total (Daniels, MacKinnon, et al. 2000).

Contrary to U.S. and Asian countries, cases of *V. parahaemolyticus* associated diseases

were rarely reported in European countries except for a few occasional outbreaks in Spain and France (Molero et al. 1989; Robert-Pillot et al. 2004; Su and Liu 2007). In addition, the outbreak case in France was caused by shrimps imported from Asia rather than domestically produced seafood (Robert-Pillot et al. 2004). Therefore, disease burdens caused from *V. parahaemolyticus* contaminated seafood are of greater concern in the U.S. and in Asian countries.

Prevalence of *V. parahaemolyticus* caused infections is globally increasing, and is likely due to global warming which has elevated the seawater temperature (Davis et al. 2017). Evidence shows that in the United States, the prevalence increased from 0.15 per million people in 1996 to 0.42 in 2010 (Newton et al. 2012). Aside from water and air temperature, there are additional risk factors that contribute to the natural abundance of *V. parahaemolyticus* like water salinity, turbidity, dissolved oxygen, organic phosphorous and nitrogen (Davis et al. 2017), which is of high public health significance for forecasting and controlling exposure to the bacterium through shellfish consumption.

Current preventive actions against *vibriosis* have been tested to be of high effectiveness and so are widely practiced. These include identifying risky beds where counts of *V. parahaemolyticus* are elevated in oysters, banning harvesting activities during the warmer months, and classifying oysters as well as other seafoods into two categories: oysters harvested in cooler seasons can be allowed for raw consumption, while those harvested in warmer months should be strongly recommended for cooking, irradiation or pasteurization since *V. parahaemolyticus* can be effectively eliminated through these treatments (Daniels, MacKinnon, et al. 2000; Su and Liu 2007).

This current study focused on shellfish harvesting waters in Washington State, which is the largest shellfish producing region in the U.S., and also has long been suffering from a high rate of *vibriosis*. Innovations of this study include spatial-temporal cluster detection to classify the most vulnerable regions, and identifying risk determinants for the infection odds of *V. parahaemolyticus* instead of abundance as previous research had explored (Davis et al. 2017), which will be useful in forecasting *vibriosis* cases and taking more

targeted prevention actions.

II METHODOLOGY

The ongoing *Vibrio* project in Washington State can be mainly divided into two procedures: shellfish sample collection analyzed for *Vibrio* spp. and *vibriosis* case trace-back, both conducted by the Washington Department of Health (WDOH). *Vibriosis* cases were confirmed by epidemiological interviews, aimed at precisely tracing back to the exact oysters consumed and their corresponding growing areas. Multiple-source infection illnesses were found occasionally when consumers had eaten various types of shellfish in a single meal. These trace-backs were censored due to the difficulty associated with successfully completing trace-backs for each food item. In total, 117 single-source illnesses

were identified from 2012 to 2016. However, single-source trace backs in 2012 were unusable because the *vibriosis* epidemiological program was just beginning and consequently only 4 trace backs were successfully completed for the entire year. Therefore, cases for year 2012 were excluded, so that only 2013-2016 recordings were included into all further analyses in this study.

Oysters were collected and tissue samples were analyzed using a most probable number polymerase chain reaction (MPN-PCR) method for three protein-directed genes: *tlh*, *tdh* and *trh* (ISSC 2015). Abiotic parameters were recorded at oyster sampling sites at the same time and place where oysters were collected from June through September each year since 2008. These parameters include *ambient air temperature* (*air*, °C), *surface water temperature* (*surface*, °C), *shore water temperature* (*shore* °C), *oyster tissue temperature* (*tissue*, °C), and *salinity* (‰). These sites where shellfish samples were collected but no infections were reported were classified as controls (n_0=1,040). For the cases (n_1=113) which were identified as the oyster growing areas where single-source illnesses were traced

back to, and where no abiotic nor biotic parameters were recorded, exact geographic locations were defined by matching oyster growing areas of the trace backs with geographical centers of all harvesting sites encompassed by the corresponding growing area. Spatial interpolations of abiotic environmental and microbial risk factors were applied to case locations through nearest neighbor re-sampling in ArcMap (ESRI 2011), taking use of the risk factor information of geographically closest control samples for dataset enrichment and further statistical analyses. Synchronized with cases, only controls samples measured during 2013-2016 were considered. All oyster growing areas where cases were collected have been mapped and marked in **Fig. 1**. Intuitively, the study districts can be divided into two regions: *Pacific Coastal Areas* and *Puget Sound Estuary*.

Basic descriptive statistics were estimated for risk factors, including central (arithmetic means, median and standard deviations) and distributional tendencies (first and third quartiles). Two-sample t-tests were used for comparing risk parameters between cases and controls directly, based on the preceding Levene's tests on variance homogeneity. Mann-

Kendall trend tests by non-parametric rank statistics were applied to directly check the variation trends across years of *vibriosis* odds (Hamed and Rao 1998; McLeod 2005; Hamed 2008), which was estimated by binary logistic regression models. Multivariate binary logistic regressions were also used for preliminary identification of risk factors and interaction effects, by means of backward stepwise selection based on criteria of likelihood ratio test (LRT), while also testing for the existence of confounding effects by checking whether the slope parameters changed more than 10% after adjusting for other risk factors. Akaike Information Criteria (AIC) were calculated and used for model selection guidance (Akaike 2011). For the purpose of explaining spatial and temporal variations and better model effectiveness hinted by AIC, time and space indicator variables (4 for years and 2 for regions) were also used in latter regression models. Principal component analysis (PCA) was applied to observe the relationships and clustering trends among all the risk factors, assisted with Pearson's coefficients of correlation, which could verify the reasonability of screening results of the potential risk determinants through stepwise logistic regression. All aforementioned statistical analyses were performed in Stata (Stata 2015) and SPSS (IBM

2016).

Spatial clustering trends of point pattern events can be evaluated by Ripley's K-function, which estimates the expected numbers of other events within a range of distances of each event, scaled by the average event spatial intensity over the whole research region (Haase 1995; Kiskowski, Hancock, and Kenworthy 2009; Waller and Gotway 2004; Dixon 2013). These scaled expectations are plotted (*y*-axis) versus distance (*x*-axis) to assess the degree to which events tend to cluster (*aka* small scale spatial dependence in point pattern data). Differences of case-control K-functions thus can be calculated as cases' K minus controls' K, which can reveal the relative clustering patterns between case and control sites, with confidence intervals estimated by Monte Carlo simulation. For large-scale (*aka* first-order) spatial variation assessment, spatial odds are estimated by intensity ratios of cases to controls to identify the high- and low-risk regions of *vibriosis*. Kernel estimating approaches for spatial odds were applied with non-parametric spatial bandwidths suggested by Diggle (Diggle and Milne 1983; Diggle 2005).

When including environmental risk factors as well as possible interaction effects as covariates, semivariograms were used as another approach for assessing small-scale spatial dependences (Waller and Gotway 2004; Garrigues et al. 2006). Semivariograms can check the spatial correlation trends of original variables (including binary variables such as case-control recordings), and of regression residuals, so as to verify whether the considered covariates have accounted for the spatial dependence of the cases and controls. Spatial statistics and geographical mappings were all accomplished in R statistical software (R 2000) with packages including *spatstat* (Adrian, Ege, and Rolf 2015), *maptools* (Bivand and Lewin-Koh 2017), and *geoR* (Ribeiro and Diggle 2016), assisted with ArcMap (ESRI 2011).

Longitudinal dependencies can be evaluated by auto-correlation coefficients of each temporal lag (Van Maanen, Nobach, and Benedict 1999), which will be frequently assisted with auto-correlation matrices aimed at visually revealing these relationships in more detail. Longitudinal variations can also be assessed by multi-level models with random intercepts

or random slopes, pooling the infection prevalence as well as risk factor associations from different groups within the same level of clusters together by Bayesian corrections based on normal distribution assumptions. For the *vibriosis* recordings, three levels can be identified: *sampling years* (2013-2016) as the highest level, *sampling regions* as the middle (*Puget Sound Estuary* or *Pacific Coastal Areas*), and *individual events* (*cases* or *controls*) as the lowest. Three-level regression models were applied to evaluate the within and across cluster divergences, indicated by intra-class correlation coefficients (ICC) (Koch 1982) of which higher values suggest stronger correlations in lower clusters. All multi-level regression analyses were accomplished in Stata.

Another approach to combine spatial and temporal exploration together is the space-time cluster detection completed in SaTScan (Norström, Pfeiffer, and Jarp 2000), regarding the epidemiological year as the additional dimension. High spatial intensity of risks can be identified from such three-dimensional analysis. It was not suitable to set months as temporal precision, since there were no records from October to May of the next year. A

spatial-temporal analysis paradigm for case-control epidemiological recordings has been graphed in **Fig. 2**.

Fig. 1 Geographical locations of oyster harvesting areas in Washington State.

Only 18 harvesting areas in total where *vibriosis* cases were reported are denoted in the map, in order to avoid unnecessary overlapping of the area labels. The harvesting areas are divided into two separated regions by land, also marked on the map. Basemap credited to ESRI, DeLorme, GEBCO, NOAA NGDC, and other contributors.

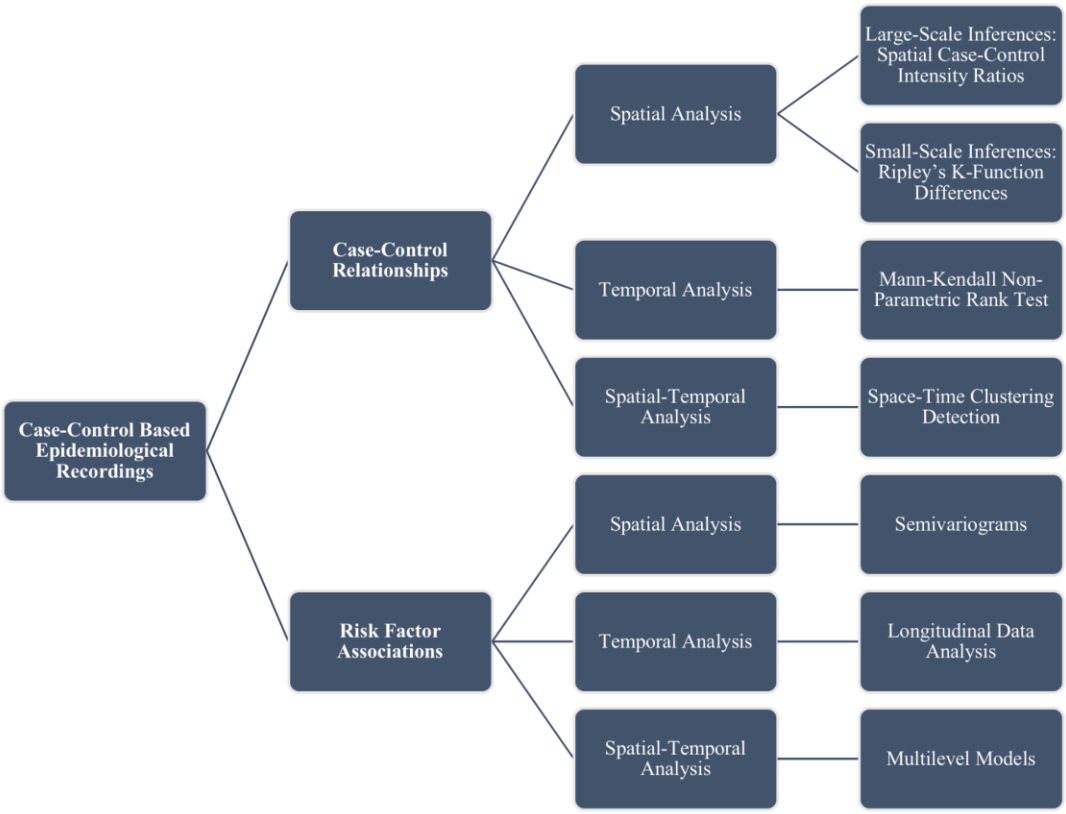

Fig. 2 Paradigm for spatial-temporal case-control analysis applied in this study.

The analysis will be conducted on two parts of datasets (the inquiry of cases and controls as shown in the second column from the left) and three types of analysis perspectives (spatial, temporal and spatial-temporal as shown in the third column). For each level and research target, corresponding methodologies have been listed in the fourth column on the right.

III RESULTS

3.1 Descriptive Summary

The originally reported records on gastro-intestinal diseases resulted from *V. parahaemolyticus* infections by contaminated seafood consumption exposure were well documented and included date and time of shellfish harvest, so that *vibriosis* odds could be estimated cumulatively for all recorded epidemiological years, and separately for each year, as listed in **Table 1**. The first row in **Table 1** presented that the overall odds of *vibriosis* is 0.11 with 95% CI of [0.09, 0.13]. The lowest *vibriosis* odds was found during year 2013 to be 0.07 (95% CI: [0.05, 0.12]), and the highest during 2016 was 0.13 (95% CI: [0.09, 0.19]).

Increasing trend was not statistically significant (p-value=0.11), which indicated that odds of *vibriosis* were not largely elevated from 2013 to 2016. Throughout the reporting period, *vibriosis* odds were estimated instead of risks. Detailed reasons will be explained in the discussion section (*Chapter* 4.1, *Page* 56). Between June and September, most cases occurred in July and August as shown in **Fig. 3**, which coincided with conclusions from previous epidemiological reports (Daniels, MacKinnon, et al. 2000). Relative spatial distribution patterns of the *vibriosis cases* and *controls* within each year are presented in **Fig. 4**.

Basic descriptive statistics are listed in **Table 2** including sample sizes, arithmetic means, standard deviations, three quartiles (25^{th}%ile, median and 75^{th}%ile), skewness and kurtosis. A note that spatial interpolation of the environmental and microbial parameters for the *case* sites was performed using a nearest neighbor re-sampling approach. Reasonability of applying nearest neighbor re-sampling method for data enrichment will be discussed further in *Chapter* 4.2 (*Page* 57). Standard deviations of the five

environmental factors were considerably smaller than their averages, hinting approximate normal distributions (skewness: [–1.83, 0.71], kurtosis: [2.84, 6.83]). On the contrary, standard deviations of the three microbial factors were tremendously larger than their averages, thus severely right-skewed distributions could be inferred considering that distances from the medians to the 75th percentiles were greater than to the 25th percentiles, which could also be verified from the positive skewness (>7.55) and rather large kurtosis (>66.2). Empirically, logarithmic transformations could be used for normalization (Sun et al. 2017; Meng et al. 2017; Chan et al. 2017), and for the convenience of further interpretations, 10 was set to be the base of log-transformation for the three microbial parameters instead of natural logarithm which was more frequently used for logistic regression. After the transformation, all three parameters were more likely to be subject to normal distribution (skewness: [–0.04, 1.00], kurtosis: [2.62, 4.88]), more suitable for parametric statistical inferences.

For normally distributed variables, Pearson's correlation coefficients were calculated

to assess the relationships, together with the scatter plots as graphed in **Fig. 5**. The four temperatures were significantly and strongly correlated with each other in positive directions, coinciding with the common sense that temperatures of ambient air, shore, surface water and oyster tissues within a region will vary similarly. Three microbial parameters were also positively related with each other, but were not correlated with the four temperature variables. *Water salinity* did not correlate with the three genetic factors, but was negatively correlated with the four temperatures, though the correlation relationships were not so strong. After linearly re-arranging the eight variables by PCA as shown from the inserted panel in **Fig. 5**, the eight factors manifested a gathering trend into three separate groups which could be labeled as temperature group, gene group, and salinity, providing a more concise description on the mutual relationships of all involved variables in this study.

In order to initially check whether these environmental and microbial factors were responsible for differentiating *vibriosis* cases and controls, two-sample t-tests were applied

for all eight risk factors by case and control groups as listed in **Table 3**. Statistically significant differences between cases and controls were observed for *surface water temperature*, *water salinity* and *tlh* MPN, indicating that these three factors should contribute to the occurrence of infection. However, t-tests are not able to account for any confounding effects, and so may provide a biased estimate of potential risk contributions. A more suitable way to observe the risk associations is via logistic regression, which will be discussed in *Chapter* 3.2 (*Page* 29).

There are other factors that could be responsible for *vibriosis*, such as the size of harvested oysters, harvesting, storing and cooling method. Smaller sized oysters were more often reported to cause *vibriosis* cases (85.5%) than larger ones, and less cases were reported when harvested by dredging (9.7%), stored in wet surroundings (20.0%) and cooled using ice or gel ices (36.1%) according to the recorded cases for year 2016 (n=32). However, the sample sizes were too small and no oyster treatment information was available for the controls to formally test this hypothesis.

Table 1 Estimated *vibriosis* odds.

Year	*n*	Cases	Odds	95% CI
all	1153	113	0.11	[0.09, 0.13]
2013	280	19	0.07	[0.05, 0.12]
2014	297	33	0.13	[0.09, 0.18]
2015	296	29	0.11	[0.07, 0.16]
2016	280	32	0.13	[0.09, 0.19]

Table 2 Descriptive statistics of abiotic and biotic parameters.

Risk Factors	*n*	Mean†	S.D. ‡	Percentiles			Distribution	
				25th	50th	75th	Skewness	Kurtosis
ambient air temperature(°C)	1153	18.5	4.3	15.6	18.0	21.0	0.71	3.93
surface water temperature (°C)	1146	19.1	2.7	17.3	18.9	20.7	0.23	3.39
shore water temperature (°C)	1146	20.1	3.5	17.7	19.7	22.1	0.48	3.20
tissue temperature (°C)	1145	21.5	5.3	17.4	20.7	25.0	0.45	2.84
water salinity (‰)	1141	25.8	6.1	25.0	27.4	29.0	–1.83	6.83
tlh (MPN)	1153	2682§	11586	15.0	93.0	750	7.55	66.2
tdh (MPN)	1153	103	3242	0.1	0.1	0.7	33.8	1147
trh (MPN)	873	274	4128	0.9	4.3	23.0	23.3	592
log_{10}-transformed tlh (*log_{10}*MPN)	1153	2.0	1.2	1.2	2.0	2.9	–0.04	2.62
log_{10}-transformed tdh (*log_{10}*MPN)	1153	–0.7	0.9	–1.0	–1.0	–0.1	1.00	4.88
log_{10}-transformed trh (*log_{10}*MPN)	873	0.6	1.0	–0.04	0.6	1.4	0.42	3.13

† Arithmetic means were calculated.

‡ S.D.: Standard Deviation.

§ The statistics including mean, standard deviation and three percentiles were rounded to 0.1 if the values were lower than 100. Otherwise, the values were kept in integers. **Table 3** followed the same settings. For skewness and kurtosis, two decimals were kept at most.

Table 3 Two-sample t-test of eight parameters between cases and controls.
Based on the Levene's homogeneity test results, equal or unequal variance assumed two-sample t-tests were selected accordingly.

Risk Factors	Cases			Controls			Levene's test	t-test
	n	Mean	S.D.	*n*	Mean	S.D.	*p*-value	*p*-value
ambient air temperature(°C)	113	19.0	4.3	1040	18.5	4.3	0.94	0.19
surface water temperature (°C)	110	18.6	2.8	1036	19.2	2.7	0.48	0.02
shore water temperature (°C)	110	19.9	3.4	1036	20.1	3.5	0.63	0.59
tissue temperature (°C)	110	21.4	5.4	1035	21.5	5.3	0.75	0.87
water salinity (‰)	110	27.1	5.5	1031	25.6	6.1	0.17	0.02
tlh (MPN)	113	1196	5063	1040	2843	12075	<0.001	0.007
tdh (MPN)	113	1.6	5.0	1040	115	3413	<0.001	0.29
trh (MPN)	94	32.8	76.0	779	304	4369	<0.001	0.08
log_{10}*-transformed tlh* (log_{10}MPN)	113	1.7	1.2	1040	2.0	1.2	0.60	0.002
log_{10}*-transformed tdh* (log_{10}MPN)	113	–0.6	0.8	1040	–0.7	1.0	0.01	0.13
log_{10}*-transformed trh* (log_{10}MPN)	94	0.6	0.9	779	0.6	1.1	0.11	0.78

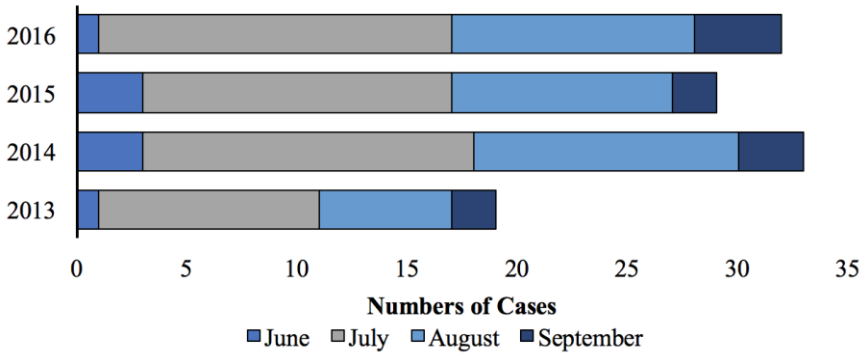

Fig. 3 Month distribution of *vibriosis* cases.

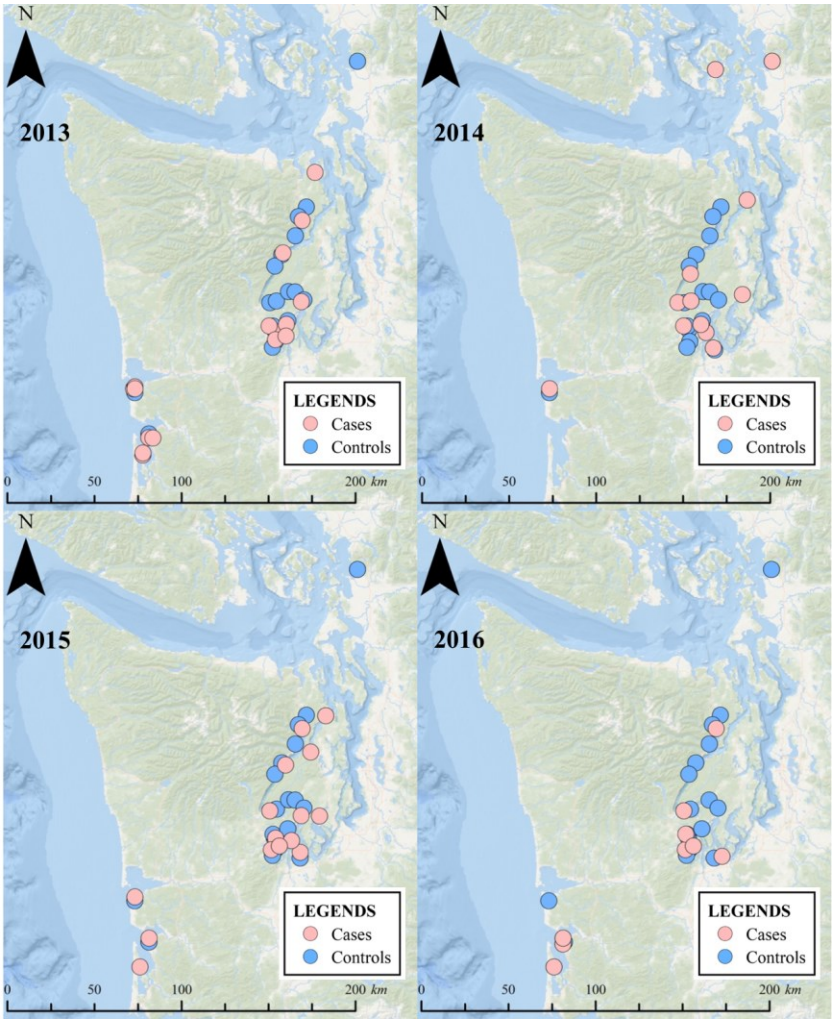

Fig. 4 *Vibriosis case* and *control* sites in Washington State across years 2013-2016.

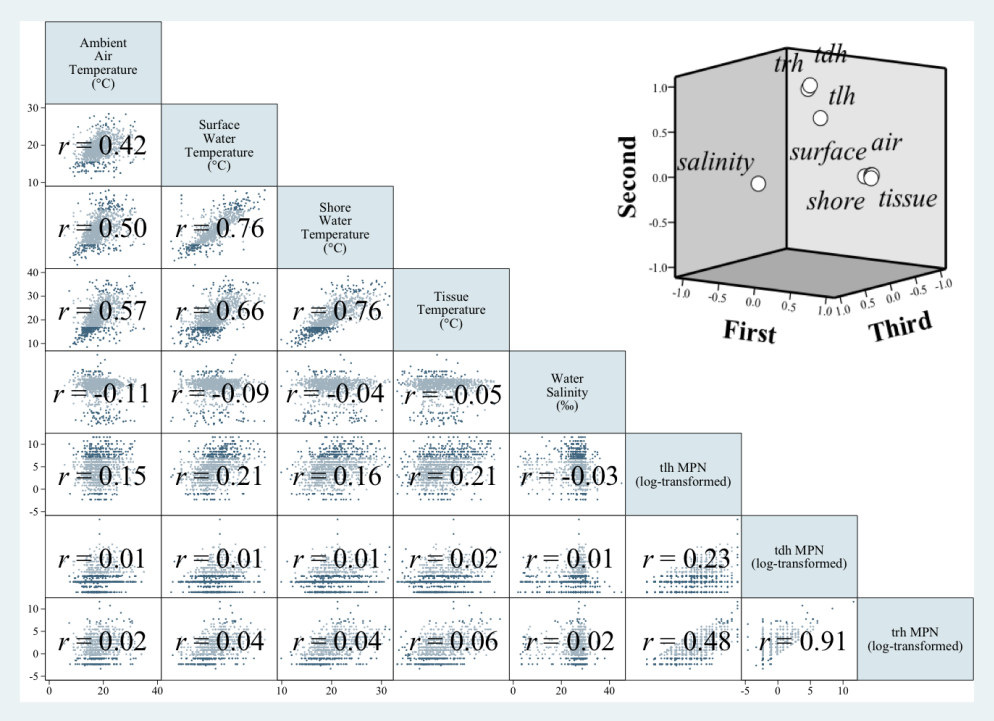

Fig. 5 Correlation coefficients matrix and factor loading graph for 8 risk factors. Coefficients of correlation were calculated by Pearson's ρ, with scatter plots added beneath each coefficient for presenting the correlations. Factor loading plot by PCA-based dimension reduction was inserted to reveal clustering relationships of the eight risk factors.

3.2 Risk Factor Association

Crude and adjusted odds ratios of all eight abiotic and biotic risk factors were estimated as summarized in **Table 4**. Before adjustment, there were three factors detected to be significantly associated with *vibriosis* odds, of which two factors contributed negatively to the odds: *surface water temperature* (°C, OR=0.92) and *tlh* (log_{10}MPN, OR=0.78), while one was positively associated: *water salinity* (‰, OR=1.05). After adjusting for other risk factors by including all the eight factors into a multivariate binary logistic regression model, four factors were statistically significant: *ambient air temperature* (°C, OR=1.06), *water salinity* (‰, OR=1.06), *tlh* (log_{10}MPN, OR=0.59), and *trh* (log_{10}MPN, OR=1.54). Though significant risk factors as well as their statistical significances were changed, no confounding effects were detected for the five environmental factors since all the odds ratios were changed less than 10% compared to the unadjusted relative odds as listed in the last column of **Table 4**.

However, confounding effects were observed for the three microbial factors, hinting

that certain factors confounded the relationship between genetic parameters and *vibriosis* odds. The fact that temperatures and water salinity were not correlated with the microbial parameters (**Fig. 5**, *Page* 28) suggested that numbers of *tlh*, *tdh* and *trh* confounded each other. Also, it was unexpected that the gene *tdh* had no substantial association with the odds of *vibriosis*, since the presence of *tdh* indicates the secretion of TDH pathogenic proteins as has been stated in the introduction section (*Page* 2). In order to explain this unexpected finding, regression models were further developed and optimized, as discussed later in this section.

Since *log*-transformed *tlh* and *trh* showed significant odds contributions as opposed to the raw values, *log*-transformed values were kept using in further analyses, and more discussions on the reasonability is included in *Chapter* 4.3 (*Page* 60).

After a series of optimization efforts as listed in **Table 5**, a credible multivariate model to assess risk associations was developed as summarized in the left part of **Table 6**, considering statistical significances of individual variables, AICs, and model

interpretations together. This model (**Model E**) was defined as the best optimized model after including interaction terms during stepwise variable selecting, but no spatial or longitudinal indicator variables were considered. **Model E** acted as a general summary of risk associations without any space-time stratifications. In order to take space-time differences into consideration, indicator variables of years and regions were attempted to be included into regression models. Whereas, only region indicators were finally kept (**Model G** marked in **Table 5**), indicating no deviations across years, which coincided with the non-elevating trend concluded from *Chapter* 3.1 (*Page* 20). On average, *vibriosis* odds in Pacific Coastal Areas were 6.9 times higher than Puget Sound Estuary, holding all involved risk factors constant. This model was defined the "Best Fitting Model" for its lowest AIC among all the attempted models. To further include time into **Model G**, a two-level mixed random intercept model clustering years (**Ultimate Model**) was developed and is summarized in the right part of **Table 5**. Since no divergences were observed in strengths of risk associations across years or regions, ransom slope models were not used. More details will be discussed in *Chapter* 4.4 (*Page* 61).

Six risk factors together with two interaction terms were finally included in the **Ultimate Model** after scrupulous optimizations: *ambient air temperature*, *surface water temperature*, *water salinity*, *tlh*, *tdh*, *trh*, and two *interaction terms*: between *surface water temperature* and *water salinity*, and between *tlh* and *tdh* (**Table 6**). This model estimates that *vibriosis* odds would be elevated by 9% due to each additional Celsius increase of *ambient air temperature* when holding all other risk factors constant. Similarly, *vibriosis* odds would be increased by 75% with each additional Celsius degree of *surface water temperature* when *water salinity* was 0‰, or due to every incremental milligram of *water salinity* when *surface water temperature* was 0°C. Besides, there was an antagonistic effect observed between *surface water temperature* and *water salinity*, such that *vibriosis* odds contributed from *surface water temperature* would be compromised by around 2% with each additional milligram of *water salinity*, and vice versa. For the microbial factors, *vibriosis* odds would be increased by 19% when *trh* MPN was amplified by an order of magnitude, would be 3.74 times higher when *tdh* MPN increased by an order of magnitude given that *tlh* MPN equaled 1, but would decrease by 40% when *tlh* MPN increased by an

order of magnitude given that *trh* MPN equaled 1. Another interaction effect was observed between *tlh* and *tdh*, such that contributions from *tdh* MPN could be compromised by 38% for every order increase of magnitude on *tlh* MPN, and *vibriosis* odds reductions from *tlh* MPN would be enlarged by 58% for every order increase of magnitude on *tdh* MPN.

Table 4 Crude and adjusted odds ratios of risk factors.

Crude odds ratios (OR) of single-source traced-back illnesses were estimated by univariate binary logistic regression models and adjusted ORs were estimated by multivariate binary logistic regression without taking any interaction terms (effect modification terms) or indicator covariates into consideration as shown below. Proportions of the changes in ORs were listed in the last 2 columns to assess existing of confounding effects.

Models for crude ORs:

$logit\ p = \beta_0 + \beta_1 \cdot air + \varepsilon$, etc.

Model for adjusted ORs:

$logit\ p = \beta_0 + \beta_1 \cdot air + \beta_2 \cdot surface + \beta_3 \cdot shore + \beta_4 \cdot tissue + \beta_5 \cdot salinity + \beta_6 \cdot tlh + \beta_7 \cdot tdh + \beta_8 \cdot trh + \varepsilon.$

Risk Factors	**Unadjusted**			**Adjusted**			**Changes (%)**
	OR	*p*-value	95% CI	OR	*p*-value	95% CI	
ambient air temperature (°C)	1.03	0.19	[0.99, 1.08]	1.06	0.03	[1.01, 1.12]	+3.36
surface water temperature (°C)	0.92	0.02	[0.85, 0.99]	0.92	0.16	[0.81, 1.04]	+0.02
shore water temperature (°C)	0.98	0.59	[0.93, 1.04]	1.05	0.40	[0.94, 1.16]	+6.20
tissue temperature (°C)	1.00	0.87	[0.96, 1.03]	0.98	0.64	[0.92, 1.05]	–1.28
water salinity (‰)	1.05	0.02	[1.01, 1.09]	1.06	0.02	[1.01, 1.12]	+1.20
tlh (MPN)	1.00	0.19	[1.00, 1.00]				
tdh (MPN)	0.99	0.28	[0.97, 1.01]				
trh (MPN)	1.00	0.27	[1.00, 1.00]				
log_{10}*-transformed tlh* (log_{10}MPN)	0.78	<0.001	[0.67, 0.91]	0.59	<0.001	[0.44, 0.80]	–23.8
log_{10}*-transformed tdh* (log_{10}MPN)	1.14	0.19	[0.94, 1.39]	1.01	0.94	[0.72, 1.44]	–11.1
log_{10}*-transformed trh* (log_{10}MPN)	0.97	0.78	[0.79, 1.19]	1.54	0.03	[1.05, 2.26]	+58.4

Table 5 AICs of all attempted binary logistic regression models.

Regions and years included into the regression models were both set to be indicator variables (dummy variables). **Model A** was named as the "**Simplest Model**" where no any covariates included; **Model E** was defined as the "**Optimized Model**" involving interaction terms after stepwise backward covariate selection; **Model G** was defined as the "**Best Fitting Model**" due to the lowest AIC. More attempts were made for selecting the final model but were not listed due to space limitation.

Model	Covariates	AIC
A	**(null)**	**741.48**
B	*air surface shore tissue salinity tlh tdh trh*	567.93
C	*air surface shore tissue salinity tlh tdh trh surface×salinity*	561.62
D	*air surface shore tissue salinity tlh tdh trh surface×salinity tlh×tdh*	552.41
E	***air surface salinity tlh tdh trh surface×salinity tlh×tdh***	**549.06**
F	*air surface salinity tlh tdh trh surface×salinity tlh×tdh i.year*	552.20
G	***air surface salinity tlh tdh trh surface×salinity tlh×tdh i.region***	**524.67**
H	*air surface salinity tlh tdh trh surface×salinity tlh×tdh i.year i.region*	527.59

Table 6 Estimated odds ratios of optimized regression and multilevel models.

Multivariate regression model represented **Model E** as summarized in **Table 5**. **Ultimate Model** referred to the two-level year-clustering random intercept mixed model developed on **Model G**, as written below. Here *i* represented the year clusters in the model. Note that the large-scale parameter estimates were identical for **Model G** and **Ultimate Model**, and so are presented in the same column.

$logit\ p = (\beta_0 + b_{0i}) + \beta_1 \cdot air + \beta_2 \cdot surface + \beta_3 \cdot salinity + \beta_4 \cdot tlh + \beta_5 \cdot tdh + \beta_6 \cdot trh + \beta_7 \cdot surface \times salinity + \beta_8 \cdot tlh \times tdh + \varepsilon.$

$b_{0i} \sim N(0, \sigma^2)$, $\varepsilon \sim N(0, \pi^2/3)$, $\sigma = 6.57 \times 10^{-7}$.

Risk Factors	**Model E**			**Model G and Ultimate Model**		
	OR	*p*-value	95% CI	OR	*p*-value	95% CI
i.region				6.90	<0.001	[3.40, 14.0]
ambient air temperature (°C)	1.07	0.01	[1.02, 1.13]	1.09	0.002	[1.03, 1.15]
surface water temperature (°C)	1.62	0.01	[1.10, 2.38]	1.75	0.005	[1.18, 2.59]
water salinity (‰)	1.62	0.001	[1.20, 2.17]	1.75	<0.001	[1.29, 2.37]
log$_{10}$*-transformed tlh* (*log*$_{10}$MPN)	0.46	<0.001	[0.32, 0.65]	0.60	0.006	[0.42, 0.86]
log$_{10}$*-transformed tdh* (*log*$_{10}$MPN)	3.70	0.003	[1.54, 8.87]	3.74	0.004	[1.53, 9.13]
log$_{10}$*-transformed trh* (*log*$_{10}$MPN)	1.51	0.04	[1.03, 2.21]	1.19	0.39	[0.80, 1.77]
surface×*salinity interaction term*	0.98	0.003	[0.97, 0.99]	0.98	0.001	[0.96, 0.99]
tlh×*tdh interaction term*	0.59	0.003	[0.42, 0.83]	0.62	0.005	[0.44, 0.87]

3.3 Clustering Divergences

Divergences in strengths of risk associations were checked across four epidemiological years and two study regions by random slope two-level mixed models clustering years or regions as written below, and the result revealed non-existences of such deviations based on the very low variances of each random effect term, as also presented.

Random slope two-level mixed model clustering years (*i for 4 years*):

$$\begin{aligned} logit\ p &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\cdot air + (\beta_2 + b_{2i})\cdot surface + (\beta_3 + b_{3i})\cdot salinity + (\beta_4 + b_{4i})\cdot tlh \\ &+ (\beta_5 + b_{5i})\cdot tdh + (\beta_6 + b_{6i})\cdot trh + (\beta_7 + b_{7i})\cdot surface \times salinity + (\beta_8 + b_{8i})\cdot tlh \times tdh. \end{aligned}$$

$$\beta_0 = -13.3, \quad b_{0i} \sim N(0, \sigma_0^2), \quad \sigma_0^2 = 1.62\times10^{-18};$$

$$\beta_1 = 0.067, \quad b_{1i} \sim N(0, \sigma_1^2), \quad \sigma_1^2 = 4.21\times10^{-18};$$

$$\beta_2 = 0.481, \quad b_{2i} \sim N(0, \sigma_2^2), \quad \sigma_2^2 = 1.16\times10^{-28};$$

$$\beta_3 = 0.480, \quad b_{3i} \sim N(0, \sigma_3^2), \quad \sigma_3^2 = 7.61\times10^{-18};$$

$$\beta_4 = -0.78, \quad b_{4i} \sim N(0, \sigma_4^2), \quad \sigma_4^2 = 4.48\times10^{-23};$$

$$\beta_5 = 1.308, \quad b_{5i} \sim N(0, \sigma_5^2), \quad \sigma_5^2 = 2.37\times10^{-19};$$

$$\beta_6 = 0.412, \quad b_{6i} \sim N(0, \sigma_6^2), \quad \sigma_6^2 = 1.53\times10^{-15};$$

$$\beta_7 = -0.02, \quad b_{7i} \sim N(0, \sigma_7^2), \quad \sigma_7^2 = 3.47\times10^{-25};$$

$$\beta_8 = -0.53, \quad b_{8i} \sim N(0, \sigma_8^2), \quad \sigma_8^2 = 8.36\times10^{-18}.$$

Random slope two-level mixed model clustering regions (*j for 2 regions*):

$$\begin{aligned} \textit{logit } p \quad &= (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})\cdot \textit{air} + (\beta_2 + b_{2j})\cdot \textit{surface} + (\beta_3 + b_{3j})\cdot \textit{salinity} + (\beta_4 + b_{4j})\cdot \textit{tlh} \\ &+ (\beta_5 + b_{5j})\cdot \textit{tdh} + (\beta_6 + b_{6j})\cdot \textit{trh} + (\beta_7 + b_{7j})\cdot \textit{surface}\times\textit{salinity} + (\beta_8 + b_{8j})\cdot \textit{tlh}\times\textit{tdh}. \end{aligned}$$

$$\beta_0 = -14.6, \qquad b_{0j} \sim N(0, \sigma_0^2), \quad \sigma_0^2 = 0.85\times10^{-00};$$

$$\beta_1 = 0.051, \qquad b_{1j} \sim N(0, \sigma_1^2), \quad \sigma_1^2 = 1.51\times10^{-14};$$

$$\beta_2 = -0.03, \qquad b_{2j} \sim N(0, \sigma_2^2), \quad \sigma_2^2 = 3.25\times10^{-03};$$

$$\beta_3 = 0.053, \qquad b_{3j} \sim N(0, \sigma_3^2), \quad \sigma_3^2 = 3.46\times10^{-16};$$

$$\beta_4 = -0.11, \qquad b_{4j} \sim N(0, \sigma_4^2), \quad \sigma_4^2 = 9.04\times10^{-21};$$

$$\beta_5 = 0.181, \qquad b_{5j} \sim N(0, \sigma_5^2), \quad \sigma_5^2 = 4.27\times10^{-24};$$

$$\beta_6 = -6.28\times10^{-3}, \quad b_{6j} \sim N(0, \sigma_6^2), \quad \sigma_6^2 = 6.32\times10^{-23};$$

$$\beta_7 = 7.84\times10^{-4}, \quad b_{7j} \sim N(0, \sigma_7^2), \quad \sigma_7^2 = 1.42\times10^{-19};$$

$$\beta_8 = 0.031, \qquad b_{8j} \sim N(0, \sigma_8^2), \quad \sigma_8^2 = 5.25\times10^{-23}.$$

In a nutshell, it could be concluded that strengths of risk associations between illnesses and the screened-out risk factors were consistent across both years and regions, and therefore there was no need to establish random effect models. Three-level random intercept mixed models were first attempted for general inferences through variance shrinkage, but variations across regions were much greater than variations across years (ICC=2.19×10^{-22}), so that it was meaningless to apply three-level models, thus the two-level random intercept mixed model clustering years were preferred. Also, although region clustering was used for assessing deviations of risk association strengths in different research regions, it was still not so appropriate to cluster the two regions in the **Ultimate Model** for odds ratio inferences, because the number of clusters for regions was so small,

which could impair the inference credibility.

In addition, for the purpose of including space variation factors, indicator variables for the two regions were added into the **Ultimate Model**, which could also be regarded as year clustering mixed model developed on **Model G**. From the two-level **Ultimate Model**, variations of *vibriosis* odds across years far exceeded the variations within years (ICC=1.31×10^{-11}), suggesting ignorable divergences across years, which could also help to explain the fact that adding the year indicator variable increased the AIC (**Table 5**), since year factor could not contribute more to explain the variations of *vibriosis* odds, but would only decrease the efficiency of the model. It is worth mentioning that no differences were observed on the odds ratios as well as 95% confidence intervals between **Model G** and the **Ultimate Model**, which also verified the consistency of *vibriosis* odds across years.

3.4 Sensitivity Analysis

Since *trh* showed significant contribution to the *vibriosis* odds while there were only 873 samples collected, which was smaller than all the other risk parameters (n=1,153) owing to the fact that *trh* began to be collected since year 2014, sensitivity analysis was conducted to check the stableness of the statistical inferences, as shown in **Table 7**.

For the five environmental factors, there were no considerable changes observed either from the arithmetic means of the variables (<2.52%), or the odds ratios estimated by univariate logistic regressions (<1.90%), before and after exclusion of recordings at 2013. Since *log*-transformed *tlh* and *tdh* MPN were taken into analyses rather than the raw values in order to exclude unreasonable influences from extreme values, sensitivity analysis was correspondingly conducted only on the normalized values. Though discrepancies in means (<5.71%) and odds ratios (<4.39%) for the two microbial parameters were both slightly greater than the environmental factors, changes were still negligible, so that such deficiency in sampling would not substantially bias the statistical inferences.

Table 7 Sensitivity analysis for 7 risk factors before and after censoring year 2013.
Odds ratios were estimated by univariate regression models.

Risk Factors	*n*		Arithmetic Mean			Odds Ratio		
	Raw	Censored	Raw	Censored	Change (%)	Raw	Censored	Change (%)
ambient air temperature (°C)	1153	873	18.5	19.0	+2.52	1.03	1.04	+0.97
surface water temperature (°C)	1146	866	19.1	19.2	+0.47	0.92	0.93	+1.09
shore water temperature (°C)	1146	866	20.1	20.3	+1.13	0.98	0.99	+1.02
tissue temperature (°C)	1145	866	21.5	21.8	+1.47	1.00	0.99	–1.00
water salinity (‰)	1141	861	25.8	25.7	–0.22	1.05	1.07	+1.90
log_{10}-transformed tlh (*log_{10}*MPN)	1153	873	2.00	2.03	+1.50	0.78	0.75	–3.85
log_{10}-transformed tdh (*log_{10}*MPN)	1153	873	–0.70	–0.66	+5.71	1.14	1.19	+4.39

3.5 Spatial Variations

Geographical information of the epidemiological recordings can be used to reveal the spatial distribution as well as interactions of the *vibriosis* events. Ripley's case-control K-function differences were calculated and demonstrated the relative clustering tendencies between the *cases* and *controls* as shown in **Fig. 6**. Across all recordings, *control* sites were observed to be more spatially aggregated than the *cases* at inter-point relative distances less than 100 *km*. However, the relative clustering trend was reversed at larger ranges with the statistical significance increasing as distance increases, suggesting that infection *cases* were clustering only in certain districts. Statistical significances had been defined by the situations where the U-shaped black curves, which represent the differences between case and control K-functions at different distances, exceeded the boundaries of two red curves, which indicate the Monte Carlo simulated confidence intervals. Similar patterns were also observed for each individual year with only slight discrepancies in the turning points of the U-shaped curves.

After confirming the spatial clustering divergences of *case* and *control* sites, specific locations of high-risk clusters were detected by geographical screening through mapping of spatial intensity ratios **Fig. 7**. In accordance with the observed trends in the K-functions, higher ratios were concentrated mainly in two regions across all years: Pacific Coastal Areas (including *Grays Harbor*, *Stony Point*, *Bay Center* and *Nahcotta*), and southern regions at Puget Sound Estuary (mainly including the *Hood Canal* series and *Totten Inlet*), which are classified in **Fig. 1**. Spatial distribution patterns of the ratios were not consistent across each year, since Pacific Coastal Areas were only of higher spatial intensity ratio in 2013 and 2016 when compared to the other two years.

Empirically, spatial clustering trends were often associated with specific factors, which were of uneven geographical distribution variations in most cases. Spatial dependences of *vibriosis cases* and *controls* in Washington State were evaluated by plotting semivariances of different inter-point relative distances as shown in the first panel of **Fig. 8**, which are likely explained by associated risk factors on which spatial dependences also occurred (**Fig.**

9). Spatial correlations were attenuated after adjusting for risk factors including *ambient air temperature*, *surface water temperature*, *water salinity*, *tlh*, *tdh* and *trh* as six predictive covariates together with two interaction terms according to the optimized regression model (**Model E** in **Table 5**). This indicates that the spatial dependences of the *vibriosis* odds were largely explained by these risk factors, since semivariances did not vary much with inter-point relative distances after adjusting spatial dependence contributions from risk factors as graphed in the right panel of **Fig. 8**.

It should be pointed out that when checking the spatial dependences after adjusting, **Model E**, the Optimized Model instead of **Model G**, the Best Fitting Model was used, since the region indicator variable is not a *real* risk factor. Dummy variables for the two study regions here were more targeted at presenting discrepancies of *vibriosis* odds in Puget Sound Estuary and Pacific Coastal Areas, rather than explaining the variations within either region. In this study, we are more interested in finding out what factors have resulted in the spatial variations of *vibriosis* odds than how the variations were distributed across our

research regions. Therefore, only contributions from meaningful risk factors were assessed.

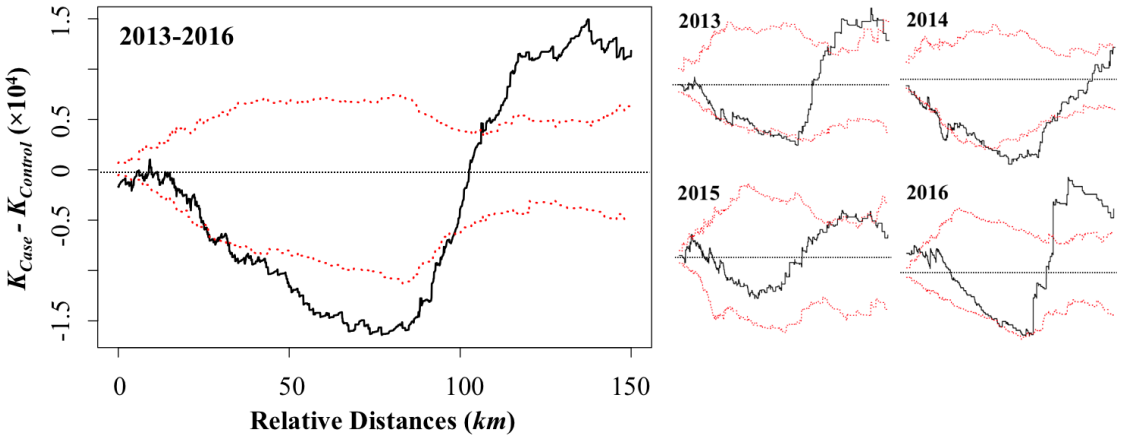

Fig. 6 Ripley's case-control K-function differences of recordings across 2013-2016. The main panel on the left side showed the case-control K-function differences of all the epidemiological recordings for half of the maximum inter-point distance. The black curve displayed absolute differences and the red curves represented the Monte Carlo simulated 95% confidence intervals. K-function differences for each individual year were plotted respectively in the right panels under the same graph settings.

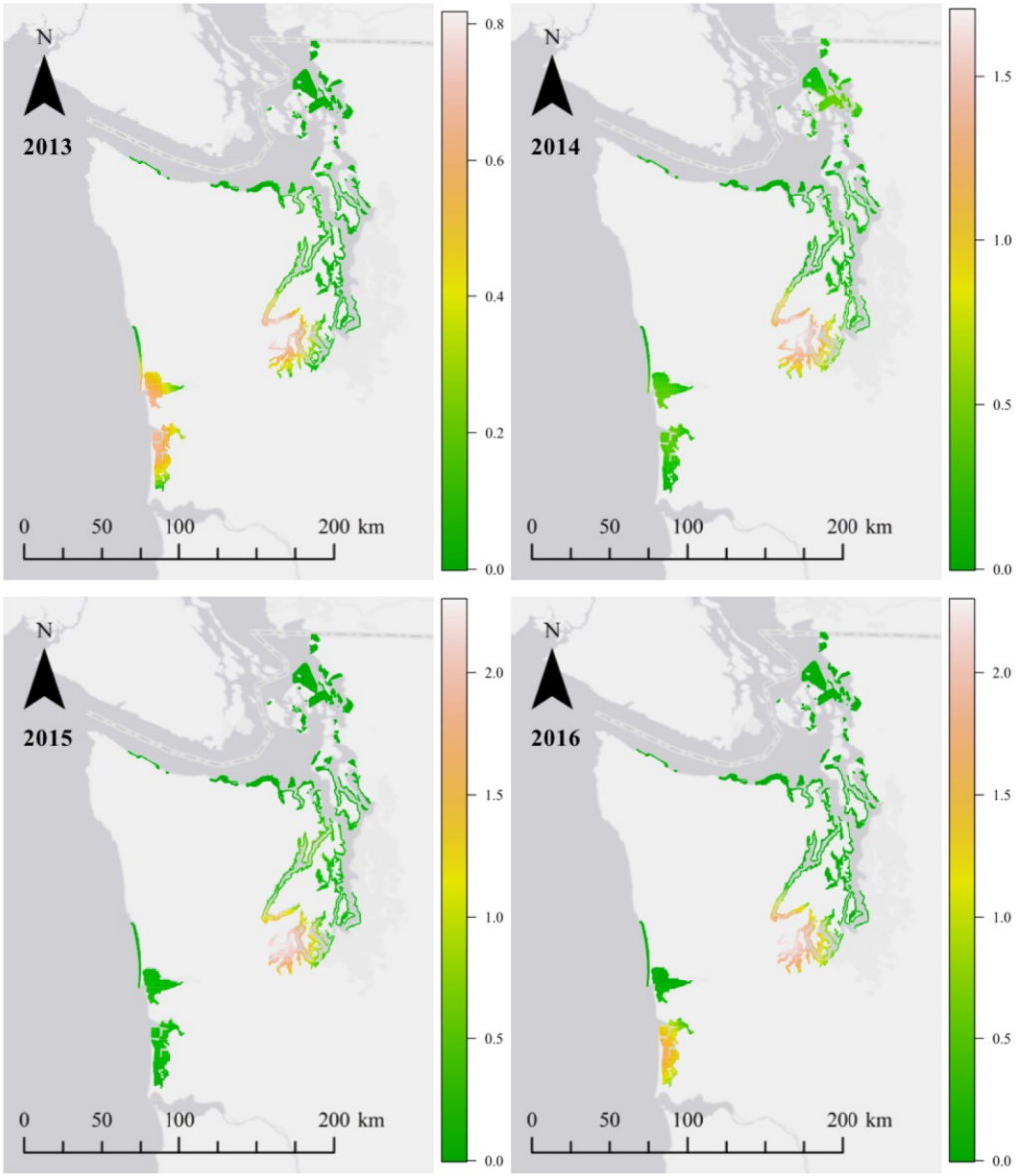

Fig. 7 Spatial intensity ratios of cases and controls across Washington State. Ratios were defined as *case* divided by *control* spatial intensities, with kernel estimations applied with 0.25 *km* bandwidths suggested by Diggle (Waller and Gotway 2004). Basemap credited to ESRI, HERE, DeLorme, MapmyIndia, © Open Street Map contributors, and the GIS user community.

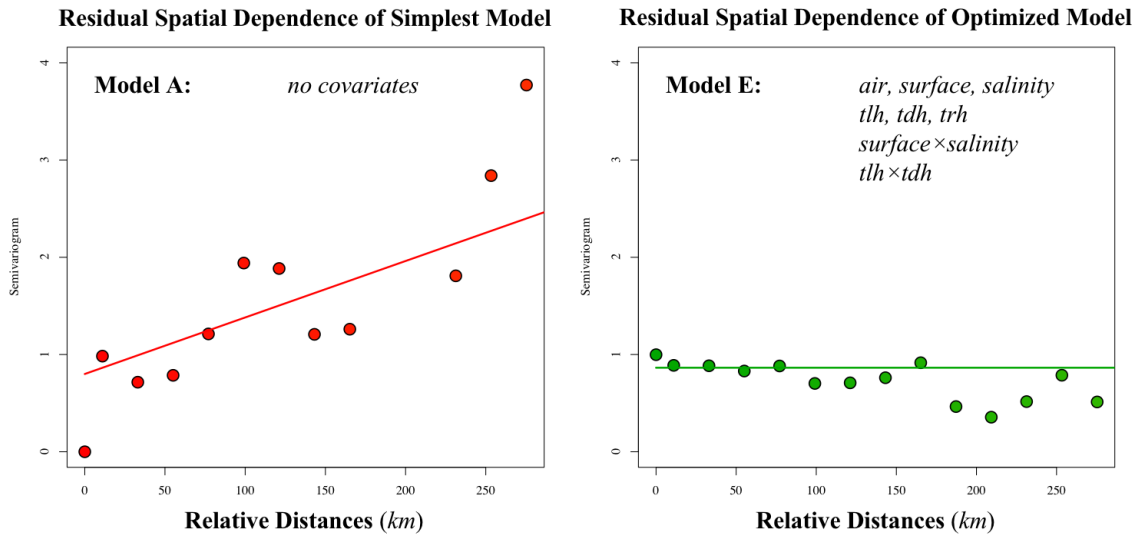

Fig. 8 Semivariograms for residuals of logistic regression models.

Inter-point semivariances were estimated for the whole extent of relative distances (maximum ~289 *km*) for two different models: **Model A**, the simplest model with no covariates included, and **Model E**, the optimized model involving stepwise selection of risk factors. Matern-class semivariogram models were applied for fitting the semivariances by weighted least squares (WLS) method. Note that the estimated sill and range parameters were so large that the semivariance function appears to be linear given the observable distances and semivariances. All the logistic regression residuals were standardized before estimating spatial dependence so that they could be used for direct comparison.

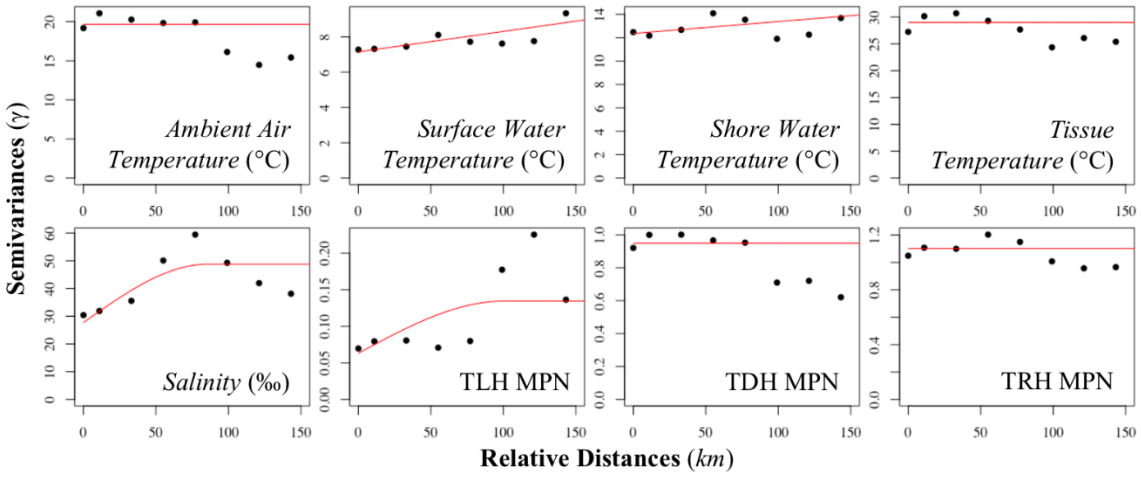

Fig. 9 Semivariograms of all 8 risk parameters.

The three microbial risk factors were log_{10}-transformed before estimating the spatial semivariances, since the raw values were of rather high variations and right-skewed distribution as had been revealed in **Table 2**. Matern-class semivariogram models were applied for fitting the semivariances by WLS as plotted in red curves.

3.6 Longitudinal Dependence

In addition to spatial dependences, temporal auto-correlations of the *vibriosis* odds were also confirmed ($|r|_{max}$ = 0.63) within each oyster growing area in the null model as presented in the lower half of **Fig. 10**. Similarly, the optimized model largely explained the longitudinal dependences, since the regression residuals were largely uncorrelated ($|r|$ < 0.07), as indicated in the upper half of **Fig. 10**. Therefore, temporal correlations are mostly attributed to the longitudinal variations of the associated risk factors. Unfortunately, temporal dependences for longitudinal regression residuals were only assessed for the latter three years, since *trh* MPN was only collected since 2014. In addition, it should be noted that although months were also recorded for all the *cases* and *controls*, time precision was still set as year so as to maintain consistency of the temporal lags.

2013	*N.A.*	*N.A.*	*N.A.*
Model A *Unadjusted* $r = 0.53$	**2014**	**Model E** *Adjusted* $r = 0.03$	**Model E** *Adjusted* $r = -0.07$
Model A *Unadjusted* $r = 0.31$	**Model A** *Unadjusted* $r = 0.19$	**2015**	**Model E** *Adjusted* $r = 0.07$
Model A *Unadjusted* $r = 0.16$	**Model A** *Unadjusted* $r = 0.02$	**Model A** *Unadjusted* $r = 0.63$	**2016**

Fig. 10 Auto-correlation matrices of year-lagged regression residuals.

Four year nodes from year 2013 to 2016 were set for longitudinal autocorrelation analysis on the unadjusted regression residuals, so that three time lags were estimated. Three years from 2014 to 2016 were assessed for adjusted residuals due to dataset deficiency. Pearson correlation coefficients were indicated in the matrices. The lower half part represented the unadjusted residuals of simple logistic regression model, while the upper half part showed the adjusted residuals of the optimized model.

3.7 Space-Time Clustering Tendency

When considering the spatial and temporal variations in odds simultaneously, space-time clusters for the extreme relative odds could be defined in three-dimensional perspectives: two dimensions denoting geographical coordinates and the remaining dimension representing the longitudinal timeline. Relative odds in this section were defined as regional *vibriosis* odds compared to overall average odds for the whole research area. To keep in accordance with conventions, and also due to the fact that risks and odds can be converted into each other, the terms "high-risk" and "low-risk" clusters were used instead of "high-odds" or "low-odds" clusters for this part of the analysis. Four high-risk space-time clusters were detected when the temporal precision was set to year, as shown in **Fig. 11**, with all the clusters extremely statistically significant (p-value < 0.001).

The cluster with the highest relative odds (OR=12.0) is at *Totten Inlet* with a *radius* of around 4.3 *km* and occurred from 2015 to 2016. The high-risk cluster with the largest area was found to be within the Pacific Coastal Areas (OR=11.0, r =13.7 *km*, 2015-2016),

covering *Bay Center*, *Stony Point* and *Nahcotta*, but not *Grays Harbor*. The remaining two high-risk clusters were respectively located around *Hood Canal 1* and *Henderson Bay* (OR=10.6 for both), occurring from 2014 to 2015.

There was only one low-risk space-time cluster identified around *Hood Canal 5*, encompassing a large range of districts with a radius of 21.0 *km* from 2015 to 2016, suggesting that in later years cases were rarely reported in these regions (OR=0.04). Not all harvesting areas in the Puget Sound Estuary fell within these four clusters, which indicated that there were still areas found to possess random occurrences of *vibriosis*. Results of the space-time cluster scanning could be useful for risk management efforts, since the risky regions and periods had been identified.

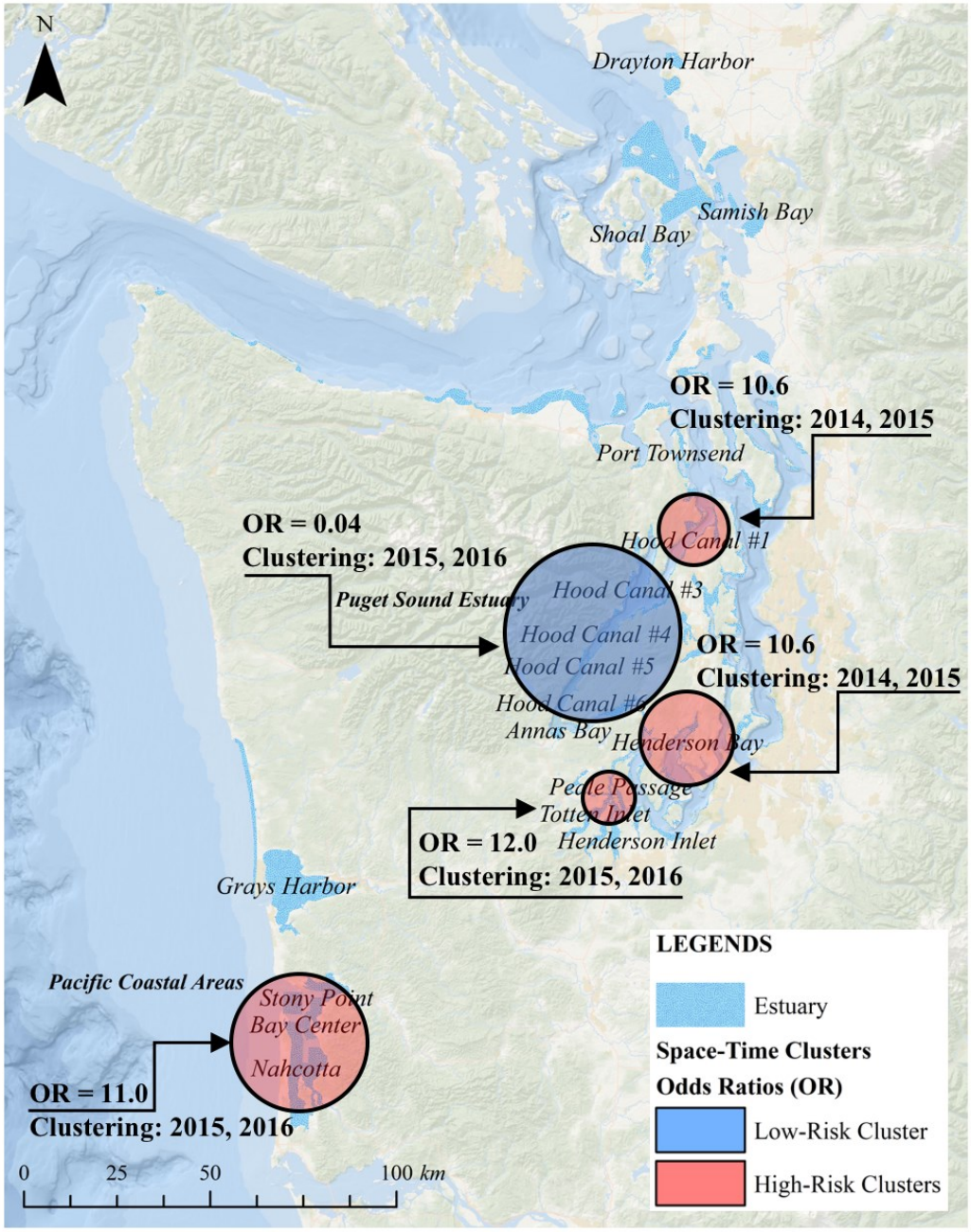

Fig. 11 Space-time clusters of extreme relative odds.
The high-risk space-time clusters are marked in red circles and low-risk marked in blue.

IV DISCUSSION

From 2013 to 2016, no significant increasing trends were observed for the *vibriosis* odds. Pacific Coastal Areas and south districts of the Puget Sound Estuary were of higher *vibriosis* odds. Three environmental and three microbial factors with two interaction effects were significantly associated with *vibriosis* odds, without any cross-year or cross-region divergences observed, and these risk factors mostly explained the spatial and temporal dependences of the odds. However, there were still several points behind the findings worth mentioning, as discussed in this chapter.

4.1 Vibriosis Odds

As far as the author is concerned, too few studies on the prevalence of *vibriosis* gastrointestinal diseases caused by consumption of raw or undercooked oysters as well as other *Vibrio*-contaminated seafoods have been conducted. Therefore, neither cross-study comparisons nor meta-analysis can currently be conducted. Unfortunately, the prevalence of *vibriosis* in Washington State could still not be inferred correctly from this study for two different reasons from perspectives of both *case* and *control* definitions.

Firstly, only successfully documented single-source trace-backs were regarded as *cases* in this study, which excluded known *vibriosis* cases that led to multi-source trace-backs. Also, we could not be totally sure that all the infections were reported and recorded, since *Vibrio*-associated gastro-intestinal disorders are self-limiting so that individuals who were infected do not bother to report their infections. These concerns are compounded by the recall bias of infected individuals in regards to which shellfish they consumed as well as unsuccessful trace-back efforts. Consequently, the number of actual *vibriosis* cases is

underestimated in these analyses. Secondly, the non-randomness of defining *vibriosis* controls by using pre-determined sample locations would also bias the inferences. The manually defined *vibriosis* controls were not equal to the whole populations at risk, as there could be a large number of residents exposed to the contaminated seafoods harvested from the same oyster growing area, but we were only taking the growing area as one *control* site. From that sense, the number of controls is not representative to the real risked populations. Therefore, using *vibriosis* odds instead of prevalence throughout these analyses was more suitable, since the real prevalence would have been drastically overestimated if we used all oyster growing areas where no infection cases were reported as controls.

4.2 Reasonability of Nearest Neighbor Re-sampling

In this study, *case* sites were located where no environmental nor microbial risk factors had been collected, and so were re-sampled by matching the *cases* with the nearest *control* sites, where parameters have been well recorded. However, there were other feasible spatial interpolation approaches to assign values to the unsampled locations in geographical

studies, such as inverse distance weighted (IDW) interpolations, using either a fixed number of variables (*variable method*), or a fixed researching radius (*radius method*) (Childs 2004; Waller and Gotway 2004).

With regard to epidemiological studies, excluding data due to information incompleteness could create biased results, although using imperfect interpolations could also lead to similar problems. For that reason, the *variable method* should be a more advantageous approach, since it is able to guarantee a relatively reasonably weighted value for all unsampled locations. Such assurances could only be accomplished for the *radius method* when the radius is defined to be sufficiently large so as to circumscribe at least one sampled site. However, using the *radius method* would impair the re-sampling efficiencies if the spatial densities of sampling locations were so low that a large radius would thus be needed to encompass at least one sampled site, since sites farther away from the unsampled points are often not so strongly correlated with the target locations to be assigned values as closer sampled sites (Waller and Gotway 2004). Semivariograms for the eight risk factors

were plotted in **Fig. 9** for the purpose of assessing spatial dependence and deciding the reasonable scanning radius for spatial interpolation. According to **Fig. 9**, the spatial variations were mostly high and stable even when the inter-point relative distances between two locations were still extremely low, indicating non-existence of spatial correlations for at least six risk parameters. Two risk factors, *salinity* and *tlh*, did display spatial dependence, thus a larger searching radius should have been applied for spatial interpolation.

However, it was not justifiable to apply different interpolation approaches within a study, especially for the same dataset. Still, it was unreasonable to use *radius method* for all eight risk factors since it would bring in uncorrelated information for the spatial independent risk factors. From that sense, *variable method* of searching one single nearest neighbor would be more beneficial. Mathematically, the IDW interpolated value with only one sampled location involved should be equal to the exact value of the referential point, just like the values of the nearest sampled sites were directly borrowed, so that the author re-named one-location *variable method* (n=1) as *nearest neighbor re-sampling*.

4.3 Normalization

The importance of using *log*-transformed *tlh*, *tdh* and *trh* MPN instead of the raw values was revealed given that all the three microbial factors seemed not to be associated with *vibriosis* odds when using the raw values in the regression analyses, while all turned out to be statistically significant after *log*-transformation. When the two-sample t-tests were conducted, significant differences were detected for *tlh* MPN (**Table 3**), but a contrary regression result showed that *tlh* MPN did not contribute anything to the odds (**Table 4**). The direct reason for this inconsistence was the heterogeneity of the *tlh* (raw values) variances for *cases* and *controls* group, since the logistic regressions assumes homogeneity of variances. However, the primary cause should be the interferences from extremely large values usually when the variables were *log*-normally distributed, since these leverage points likely conceal the actual differences and associations. Therefore, it was necessary to first normalize the extremely right-skewed variables before performing regression analyses in case associations would be masked by extreme values to be insignificant. Additionally,

normalization of the raw variables would be beneficial for correlation relationship exploration and visualization (**Fig. 5**). As the most widely used transformation approach from Box-Cox transformation series, *log*-transformation has been verified to be suitable in a wide range of studies (Sun et al. 2017; Meng et al. 2017; Balakrishnan et al. 2013).

4.4 Risk Association Interpretations

An unexpected result observed from comparing the odds ratios before and after adding interaction terms was the reversed risk association direction on *surface water temperature* (**Table 4** and **Table 6**). Without the effect modification term, the average odds ratio of the *surface water temperature* was around 0.92, which was a negative association. However, the association became positive, with an odds ratio around 1.62, after including the interaction terms with *water salinity*. A credible explanation on this reversion was the antagonistic interaction effects that increasing of *surface water temperature* would become negatively associated with *vibriosis* odds when *water salinity* was higher than 22.9‰. In total, 90.6% samples had water salinity values higher than 22.9‰, so that the averaged

odds ratio of *surface water temperature* would be in negative direction if not stratifying the *water salinity*. For *water salinity*, the odds association was positive only when the *surface water temperatures* were lower than 22.4°C. In total, 88.0% samples of *surface water temperature* were no higher than 22.4°C, so that the average odds ratio of *water salinity* was still positive without any reversion observed.

Another unexpected result was the negative association between *vibriosis* odds and *tlh* abundance, since no previous researches revealed such associations either from epidemiological recordings or pathogenic mechanisms. Although *tdh* and *trh* are able to synthesize and secrete virulent proteins, the two genes may not necessarily lead to illnesses (Davis et al. 2017). Therefore, it could be hypothesized that when numbers of *tlh* are quite large, the probability of expression on gene *tdh* and *trh* among the *V. parahaemolyticus* community would be decreased, and consequentially reduce the odds of *vibriosis* indirectly.

The third unexpected result was the lack of a statistically significant odds contribution from *tdh* MPN when no interaction term was taken into consideration. The direct reason

was the fact that *tlh* MPN was not a confounder of *tdh* MPN, but instead acted as an effect modifier. When *tlh* MPN was lower than 200 (*log_{10}-transformed tlh* MPN was lower than 2.3), then the risk association direction of *tdh* MPN would be positive; contrary, when *tlh* MPN was higher than 200, the *tdh* MPN risk association direction would turn negative. There were 56.3% samples being of *tlh* MPN lower than 200, which was quite close to 50%. Therefore, the average association of *tdh* MPN was compromised to be insignificant when not including an interaction term due to the comparable sample sizes of positive and of negative associations. This effect modification, along with the overall association with *tlh* suggests that the proportion of *tdh* relative to *tlh* in a shellfish sample may be a stronger indicator of *vibriosis* risk than each individual genetic target. Washington State shellfish harvesting waters are well-known for high *tdh* levels and this analysis provides new insight into how the relative abundance of genetic markers can indicate a higher foodborne risk for shellfish harvested in this region. Future analyses should investigate this idea further by directly applying a *tdh*:*tlh* ratio to the models used in this analysis.

Overall, higher *tdh* and *trh* MPN were positively associated with *vibriosis* odds, which was in accordance with conclusions from previous study reports (Raimondi et al. 1995; Raimondi et al. 2000; Su and Liu 2007; Miyamoto et al. 1969). No interaction terms between environmental factors and microbial factors were tested to be of significance, which is supported by the lack of correlations between these variables as shown in **Fig. 5**, indicating that environmental parameters and virulent gene numbers affected the *vibriosis* odds separately.

4.5 Spatial Clustering

The phenomena concerning spatial case-control distribution patterns such that *controls* were more clustered in smaller ranges of districts but *cases* were more aggregated at larger ranges can be attributed to lower distances between *control* sites, and also the moderate *case* sample size. For this reason, more *control* sites would be circumscribed into the scanning circles with low radius while very few cases could be included. But from larger spatial scales, *cases* were more clustered into two specific regions than the dispersedly

distributed controls, since the scanning circles with greater radius could encompass more cases so that the real clustering tendencies were no longer concealed. Such down-and-up U-shaped tendencies likely indicate the region-specific clustering of cases, and often the clusters were isolated respectively, which matched the reality that *cases* were reported mostly in southern districts of Puget Sound Estuary, and Pacific Coastal Areas these two clusters.

V IMPLICATIONS

Although gastro-intestinal diseases from *Vibrio parahaemolyticus* infections by exposure to contaminated oysters had long been confirmed as a public health concern, especially in countries and regions where habits of consuming undercooked or even raw seafoods are kept, there are still no previous studies on *vibriosis* for Washington State as far as the author is aware. Since oyster harvesting activities are proceeding in the Puget Sound Estuary and the Pacific Coastal Areas while *vibriosis* cases caused by oyster consumptions are frequently reported in recent years (McLaughlin et al. 2005; Daniels, MacKinnon, et al. 2000; DePaola et al. 2000), our research will be quite important for

evaluating infection risks in Washington State.

Even though epidemiological research reports on *vibriosis* cases have been published globally, there are still very few studies conducted on identifying environmental risk factors of *vibriosis* illnesses. Therefore, this work in will be of high research significance to identify best practices for risk management, since we have directly taken environmental temperatures and water salinity into consideration for the odds of *vibriosis*. Our study results will also be useful in *vibriosis* risk forecasting by monitoring the environmental parameters, which is less costly and easier to conduct than sampling oysters.

This research project is one of the very few epidemiological studies that uses a nearest neighbor re-sampling approach for data enrichment. Moreover, it is innovative to verify the reasonability of using nearest neighbor re-sampling method from perspectives of spatial variations by means of semivariances, which is coherent to the statistical principles and can be used in future research.

We have also discussed the importance of normalizing the genetic variables before

statistical inferences, which can direct future studies to avoid possible mistakes. Exploring the auto-correlation relationships of all the variables by PCA is also helpful in identifying interaction terms before performing regression analysis. For spatial and temporal variations and dependences, we have combined space-time clustering scanning and multilevel mixed models to screen out the most vulnerable regions and periods from perspectives of large scale inferences, and to assess similarities and differences across regions and years from perspectives of small scale inferences.

VI LIMITATIONS

Although there are significant research findings as well as a lot of analytical innovations in this study, several limitations are still unavoidable.

The primary limitation is the inability to infer *vibriosis* prevalence, as has been discussed in *Chapter* 4.1 (*Page* 56). WDOH has set a *vibriosis* risk control target that prevalence should be below 1 out of 100,000 population (Paranjpye et al. 2015), but unfortunately our current research cannot provide any information for comparison.

Secondly, we did not have any measured risk factors for the case harvesting sites and so used a nearest neighbor re-sampling approach. Although it is a quite fantastic way for

data enrichment as has been discussed in *Chapter* 4.2 (*Page* 57), bias will still be inevitable. For example, the nearest neighbor re-sampling approach will cause information loss when spatial correlations exist, like for *water salinity* and *tlh* MPN (**Fig. 9**). Another type of bias stems from the fact that using closest Euclidean distances may not be an optimal choice as the samples were collected in water environments where water channel distances may be more credible for determining the smallest distances. However, due to lack of water channel information, the water channel distances could not be estimated and so could not be used throughout this research.

Thirdly, we only considered four temperatures and water salinity as environmental risk parameters, while there should be more factors taken into our consideration like water turbidity and dissolved oxygen (Davis et al. 2017). Although the ambient air and surface temperatures together with water salinity were verified to be strongly associated with *vibriosis* odds in our study (see in *Chapter* 3.2, *Page* 29), we do not know whether such relationships were confounded by unmeasured factors. Besides, we do not have any ideas whether the pathogenic genes are interacting with these environmental factors not collected

in our study, though the environmental factors included in this study seemed not to be in relationship with microbial parameters.

Post-harvesting methods of the harvested oysters, like wet storage, ice cooling, etc., strongly affect the growth of *Vibrios* (Cook 1997; Cook and Ruple 1992; Andrews, Park, and Chen 2000; Songsaeng et al. 2010). However, we have rather limited recordings on these factors as treatment information was only collected for the cases in 2016, thus no comparisons between cases and controls can be made, and the sample sizes are too low to ensure trust any statistical inferences.

Though we also assessed the temporal variations and dependences of the *vibriosis* odds, we had only four years of epidemiological recordings, for which the sample size was too small. What's worse, *trh* MPN began to be collected since 2014, so that we only had effective recordings for three years (2014-2016), which weakened our temporal relationship explorations.

However, we are optimistic that we can improve on these deficiencies in the future, and that these concerns can also be useful to other researchers who will be conducting similar or related studies.

VII CONCLUSIONS

Gastro-intestinal illnesses resulted from *Vibrio parahaemolyticus* infections by consumption of raw or undercooked oysters were substantial in Washington State in the past several years, which needs to be addressed since it is a common disease burden frequently reported in warmer seasons. Although, fortunately, no increasing tendency of the prevalence was observed, occurrence of infections still had considerable spatial variations across different oyster growing areas, suggesting that high risks were geographically aggregated into specific regions. Such spatial divergences of risks were associated with both environmental and microbial factors, but no interaction effects were

found between these two sets of variables. Although infection risks varied drastically across regions, associations of risk factors were constant both spatially and temporally. These results suggest that risk management of *Vibriosis* can be improved by primarily harvesting oysters from areas where ambient air temperature is lower, and surface water temperature and salinity are both relatively very low or very high as the first step of risk control. Considering the fact that ambient air temperatures and surface water temperatures vary synchronically, it will be more realistic to conduct oyster harvesting activities where ambient air temperature, surface water temperature and salinity are all at a lower level. Also, identifying high *tdh* and low *tlh* most probable numbers is another preventative monitoring technique as well from perspectives of genetic parameters. Wet storage and ice cooling for the harvested oysters can likely decrease the infectious dose of *Vibrio*, which could be suggested to shellfish harvesting companies for the purpose of reducing infection possibilities by adjusting seafood treatment methods as the second step of risk control.

APPENDICES

Stata Programming Codes

For Chapter 3.1:

- Estimating the *vibriosis* odds overall and individually for years (**Table 1**, *Page* 25):

```
bys year: logistic case
logistic case
nptrend case, by (year)
```

- Providing descriptive statistics for all involved variables (**Table 2**, *Page* 25):

```
summarize air surface shore tissue salinity tlh tdh trh, detail
```

- Levene's test and two-sample t-test (**Table 3**, *Page* 26):

```
sdtest air, by (case)
ttest air, by (case)
ttest air, by (case) unequal
```

For Chapter 3.2:

- Estimating the crude and adjusted odds ratios (**Table 4**, *Page* 34):

logistic case air
logistic case air surface shore tissue salinity tlh tdh trh

- Stepwise variable selection and AIC calculation (**Table 5**, *Page* 35):

stepwise pr(0.2): logistic case air surface shore tissue salinity tlh tdh trh ss ld
estat ic

- Estimating deviations of risk association strengths across years and regions by random slope mixed models (*Page* 37):

meqrlogit case air surface salinity tlh tdh trh ss ld || year: air surface salinity tlh tdh trh ss ld, or
meqrlogit case air surface salinity tlh tdh trh ss ld || region: air surface salinity tlh tdh trh ss ld, or

- Estimating odds ratios of each risk factor by the Ultimate Model (**Table 6**, *Page* 36):

meqrlogit case i.region air surface salinity tlh tdh trh ss ld || year: , or
estat icc

For Chapter 3.6:

- Checking longitudinal dependences before and after adjustment (**Fig. 10**, *Page* 51):

logistic case i.year
predict raw, resid
autocor raw year growingarea

logistic case air surface salinity tlh tdh trh ss ld
predict adjusted, resid
autocor adjusted year growingarea

R Programming Codes

For Chapter 3.5:

- Calculating case-control K-function differences (**Fig. 6**, *Page* 47):

library(nlme); library(rpart); library(spatstat); library(sp); library(splancs);
library(maptools); library(maps); library(geoR)

setwd(""); vibriosis<-read.csv("Vibriosis.csv")

long<-vibrio$X; lat<-vibrio$Y; sites<-cbind(long,lat)
plot(sites); poly<-getpoly()

```
vibrio.cases<-vibrio[vibrio$Case==1,c("X","Y")]
vibrio.cntls<-vibrio[vibrio$Case==0,c("X","Y")]
vibrio.cases<-as.points(x=vibrio.cases[,1],y=vibrio.cases[,2])
vibrio.cntls<-as.points(x=vibrio.cntls[,1],y=vibrio.cntls[,2])


max.dist<-max(dist(cbind(vibrio$X,vibrio$Y)))
h<-seq(0,150,length=10000)

k1<-khat(vibrio.cases,poly,h)
k0<-khat(vibrio.cntls,poly,h)
kdiff<-k1-k0
env.label<-Kenv.label(vibrio.cases,vibrio.cntls,poly,nsim=99,h)
plot(h,kdiff,pch=" ",xlab="Distance (km)",ylab="K cases - K controls")
lines(h,kdiff,lwd=2,col="black")
lines(h,env.label$upper,lty=3,lwd=2,col="red")
lines(h,env.label$lower,lty=3,lwd=2,col="red")
```

- Estimating spatial ratios of *vibriosis* odds by kernel estimation (**Fig. 7**, *Page* 48):

```
kern.vibrio.case<-kernel2d(vibrio.cases,poly,h0=250,nx=100,ny=100)
polymap(poly,axes=FALSE)
image(kern.vibrio.case,add=T)
points(vibrio.cases,pch=16,col="green")

kern.vibrio.cntl<-kernel2d(vibrio.cntls,poly,h0=250,nx=100,ny=100)
polymap(poly,axes=FALSE)
image(kern.vibrio.cntl,add=T)
points(vibrio.cntls,pch=1,col="blue")

kern.ratio<-kernrat(as.points(x=vibrio.cases[,1],y=vibrio.cases[,2]),
                    as.points(x=vibrio.cntls[,1],y=vibrio.cntls[,2]),
                    poly,h1=250,h2=250,nx=100,ny=100)

polymap(poly,axes=FALSE)
image(kern.ratio,col=heat.colors(12),add=TRUE)
points(vibrio.cntls,pch=16,col="blue")
points(vibrio.cases,pch=16,col="green")
legend(locator(1),c("Case","Control"),pch=c(16,1),
       col=c("green","blue"),bty="n")

library(spam)
library(grid)
library(fields)

kern.ratio<-kernrat(as.points(x=vibrio.cases[,1],y=vibrio.cases[,2]),
                    as.points(x=vibrio.cntls[,1],y=vibrio.cntls[,2]),
                    poly,h1=250,h2=250,nx=100,ny=100)

polymap(poly,axes=FALSE,lwd=2)
image.plot(kern.ratio,col=terrain.colors(256),add=T)
```

- Assessing spatial auto-correlations before and after adjustment (**Fig. 8**, *Page* 49):

```
model.0<-glm(Case~1,family=binomial,data=vibriosis)
model.1<-glm(Case~air+surface+salinity+ss+tlh+tdh+trh+ld,family=binomial,data=vibriosis)


summary(model.0)
summary(model.1)

resids.0<-(model.0$y-model.0$fitted.values)/sqrt(model.0$fitted.values*(1-model.0$fitted.values))
resids.1<-(model.1$y-model.1$fitted.values)/sqrt(model.1$fitted.values*(1-model.1$fitted.values))

geo.0<-as.geodata(cbind(vibriosis$X,vibriosis$Y,resids.0))
geo.1<-as.geodata(cbind(vibriosis$X,vibriosis$Y,resids.1))

v.0<-variog(geo.0,max.dist=289)
v.1<-variog(geo.1,max.dist=289)

plot(v.0,xlab="Relative Distance (km)",ylab="Semivariogram",
     pch=21,ylim=c(0,4),cex=2.0, col="black", bg=heat.colors(100), lwd=1.6,
     main="Residual Spatial Dependence of Simplest Model")
v.0.wls<-variofit(v.0,ini.cov.pars=c(2,150),cov.model="matern",nugget=2,weights="cressie")
lines(v.0.wls, col=heat.colors(12), lwd=2.4)

plot(v.1,xlab="Relative Distance (km)",ylab="Semivariogram",
     pch=21,ylim=c(0,4),cex=2.0, col="black", bg=terrain.colors(100), lwd=1.6,
     main="Residual Spatial Dependence of Optimized Model")
v.1.wls<-variofit(v.1,ini.cov.pars=c(0,0),cov.model="linear",nugget=1,weights="cressie")
lines(v.1.wls, col=terrain.colors(12), lwd=2.4)
```

ACKNOWLEDGEMENT

To my academic advisor **Dr. Frank C. Curriero**, who broadened my statistical skills by a brand new perspective on spatial analysis, led me to construct research ideas and accomplish the degree requirements. Frank helped me a lot on revising my research proposal and final thesis, no matter in contents or in language. And also to **Dr. Benjamin J. Davis**, who put forward lots of detailed constructive comments on my thesis.

To **Dr. Ernst W. Spannhake**, who invited me to this best school of public health all over the world, and acted as my primary advisor. To **Dr. Marie Diener-West**, **Dr. Karen Bandeen-Roche**, **Dr. Elizabeth Colantuoni**, and **all other course instructors**, who

amplified my knowledge and elevated my research skills into an advanced level.

To **Ms. Katie Phipps**, the student affair administrator at Department of Environmental Health and Engineering, who helped me move through complicated affairs patiently and carefully for countless times.

To **Ms. Clara Hard** from Washington Department of Health, who conducted the illness traceboack, as well as the **Washington Department of Health State Laboratory** where the microbial analyses were finished, so that I can achieve this work. To the **Fulbright Program**, who sponsored me all through my oversea education experiences.

To **Bloomberg School of Public Health, Johns Hopkins University**, and **Peking University**, who helped me lay the soild foundation of environment-health association researches in my future.

To my family, who supported me all the way through thick and thin.

Also, to myself, because I never give up.

THANK YOU. You are all of great meanings to my life. What I can only do, is to make your meanings more meaningful.

BIBLIOGRAPHY

Adrian, B., R. Ege, and T. Rolf. 2015. *Spatial Point Patterns: Methodology and Applications with R* (CRC Press).

Akaike, H. 2011. 'Akaike's Information Criterion.' in, *International Encyclopedia of Statistical Science* (Springer).

Andrews, L. S., D. L. Park, and Y. P. Chen. 2000. 'Low temperature pasteurization to reduce the risk of *vibrio* infections from raw shell-stock oysters', *Food Addit Contam*, 17: 787-91.

Bag, Prasanta, Suvobroto Nandi, Rupak K Bhadra, Thandavarayan Ramamurthy, SK Bhattacharya, M Nishibuchi, T Hamabata, Shinji Yamasaki, Yoshifumi Takeda, and G Balakrish Nair. 1999. 'Clonal Diversity among Recently Emerged Strains of *Vibrio parahaemolyticus* O3:K6 Associated With Pandemic Spread', *Journal of clinical microbiology*, 37: 2354-57.

Balakrishnan, K., S. Ghosh, B. Ganguli, S. Sambandam, N. Bruce, D. F. Barnes, and K. R. Smith. 2013. 'State and national household concentrations of $PM_{2.5}$ from solid cookfuel use: results from measurements and modeling in India for estimation of the global burden of disease', *Environ Health*, 12: 77.

Bivand, R., and N. Lewin-Koh. 2017. "maptools: Tools for Reading and Handling Spatial Objects." In *R package Version 0.9-2*.

Blake, P. A., R. E. Weaver, and D. G. Hollis. 1980. 'Diseases of humans (other than cholera) caused by *vibrios*', *Annu Rev Microbiol*, 34: 341-67.

Cai, J., Y. Han, and Z. Wang. 2006. 'Isolation of *Vibrio parahaemolyticus* from abalone (Haliotis diversicolor supertexta L.) postlarvae associated with mass mortalities', *Aquaculture*, 257: 161-66.

CDC. 1998. 'Outbreak of *Vibrio parahaemolyticus* infections associated with eating raw oysters: Pacific Northwest, 1997', *Morbidity and Mortality Weekly Report*, 47: 457.

———. 1999. 'Outbreak of *Vibrio parahaemolyticus* infection associated with eating raw oysters and clams harvested from Long Island Sound - Connecticut, New Jersey, and New York, 1998', *Annals of Emergency Medicine*, 34: 679-80.

Chen, Shiyang, Shuqing Liu, and Lifang Zhang. 1991. 'Occurrence of *Vibrio parahaemolyticus* in seawater and some seafoods in the coastal area of Qingdao', *Journal of Ocean University of Qingdao*, 21: 43-51.

Childs, Colin. 2004. 'Interpolating surfaces in ArcGIS spatial analyst', *ArcUser, July-September*, 3235: 569.

Cook, D. W. 1997. 'Refrigeration of oyster shellstock: Conditions which minimize the outgrowth of *Vibrio vulnificus*', *Journal of Food Protection*, 60: 349-52.

Cook, D. W., and A. D. Ruple. 1992. 'Cold-Storage and Mild Heat-Treatment as Processing Aids to Reduce the Numbers of *Vibrio vulnificus* in Raw Oysters', *Journal of Food Protection*, 55: 985-89.

Dadisman, T. A., R. Nelson, J. R. Molenda, and H. J. Garber. 1973. '*Vibrio parahaemolyticus* Gastroenteritis in Maryland - Clinical and Epidemiologic Aspects', *Journal of Milk and Food Technology*, 36: 111-12.

Daniels, N. A., L. MacKinnon, R. Bishop, S. Altekruse, B. Ray, R. M. Hammond, S. Thompson, S. Wilson, N. H. Bean, P. M. Griffin, and L. Slutsker. 2000. '*Vibrio parahaemolyticus* infections in the United States, 1973-1998', *J Infect Dis*, 181: 1661-6.

Daniels, N. A., B. Ray, A. Easton, N. Marano, E. Kahn, A. L. McShan, 2nd, L. Del Rosario, T. Baldwin, M. A. Kingsley, N. D. Puhr, J. G. Wells, and F. J. Angulo. 2000. 'Emergence of a new *Vibrio parahaemolyticus* serotype in raw oysters: A prevention quandary', *Jama*, 284: 1541-5.

Davis, B. J. K., J. M. Jacobs, M. F. Davis, K. J. Schwab, A. DePaola, and F. C. Curriero. 2017. 'Environmental determinants of *Vibrio parahaemolyticus* in the Chesapeake Bay', *Appl Environ Microbiol*: 01147-17.

Deepanjali, A., H. S. Kumar, I. Karunasagar, and I. Karunasagar. 2005. 'Seasonal variation in abundance of total and pathogenic *Vibrio parahaemolyticus* bacteria in oysters along the southwest coast of India', *Appl Environ Microbiol*, 71: 3575-80.

DePaola, A., C. A. Kaysner, J. Bowers, and D. W. Cook. 2000. 'Environmental investigations of *Vibrio parahaemolyticus* in oysters after outbreaks in Washington, Texas, and New York (1997 and 1998)', *Appl Environ Microbiol*, 66: 4649-54.

Diggle, P. 2005. 'Applied Spatial Statistics for Public Health Data.' in (Taylor & Francis).

Diggle, P. J., and R. K. Milne. 1983. 'Bivariate Cox Processes - Some Models for Bivariate Spatial Point Patterns', *Journal of the Royal Statistical Society Series B-Methodological*, 45: 11-21.

Dixon, Philip M. 2013. 'Ripley's K function', *Encyclopedia of environmetrics*.

ESRI. 2011. "ArcGIS Desktop: Release 10." In. Redlands, CA: Environmental Systems Research Institute.

Fabbri, A., L. Falzano, C. Frank, G. Donelli, P. Matarrese, F. Raimondi, A. Fasano, and C. Fiorentini. 1999. '*Vibrio parahaemolyticus* thermostable direct hemolysin modulates cytoskeletal organization and calcium homeostasis in intestinal cultured cells', *Infect Immun*, 67: 1139-48.

Fujino, T, R Sakazaki, and K Tamura. 1974. 'Designation of the type strain of *Vibrio parahaemolyticus* and description of 200 strains of the species', *International Journal of Systematic and Evolutionary Microbiology*, 24: 447-49.

Garrigues, S, D Allard, F Baret, and M Weiss. 2006. 'Quantifying spatial heterogeneity at the landscape scale using variogram models', *Remote Sensing of Environment*, 103: 81-96.

Haase, Peter. 1995. 'Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction', *Journal of vegetation science*, 6: 575-82.

Hamed, K. H. 2008. 'Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis', *Journal of hydrology*, 349: 350-63.

Hamed, K. H., and A. R. Rao. 1998. 'A modified Mann-Kendall trend test for autocorrelated data', *Journal of hydrology*, 204: 182-96.

Hara-Kudo, Y., T. Nishina, H. Nakagawa, H. Konuma, J. Hasegawa, and S. Kumagai. 2001. 'Improved method for detection of *Vibrio parahaemolyticus* in seafood', *Appl Environ Microbiol*, 67: 5819-23.

Hlady, W. G., and K. C. Klontz. 1996. 'The epidemiology of *Vibrio* infections in Florida, 1981-1993', *J Infect Dis*, 173: 1176-83.

Honda, S., I. Goto, I. Minematsu, N. Ikeda, N. Asano, M. Ishibashi, Y. Kinoshita, M. Nishibuchi, T. Honda, and T. Miwatani. 1987. 'Gastroenteritis Due to Kanagawa Negative *Vibrio parahaemolyticus*', *Lancet*, 1: 331-32.

Honda, T., Y. X. Ni, and T. Miwatani. 1988. 'Purification and characterization of a hemolysin produced by a clinical isolate of Kanagawa phenomenon-negative *Vibrio parahaemolyticus* and related to the thermostable direct hemolysin', *Infect Immun*, 56: 961-5.

Honda, T., Y. X. Ni, T. Miwatani, T. Adachi, and J. Kim. 1992. 'The Thermostable Direct Hemolysin of *Vibrio parahaemolyticus* Is a Pore-Forming Toxin', *Canadian journal of microbiology*, 38: 1175-80.

IBM. 2016. "SPSS Statistics: Release 24." In. Armonk, NY: IBM Corporation.

ISSC. 2015. 'Method: *Vibrio parahaemolyticus* enumeration and detection through MPN and real-time PCR', Accessed January 9, 2015. http://www.issc.org/issc-task-force-i-proposals.

Kaneko, T., and R. R. Colwell. 1973. 'Ecology of *Vibrio parahaemolyticus* in Chesapeake Bay', *J Bacteriol*, 113: 24-32.

Kaper, J. B., R. K. Campen, R. J. Seidler, M. M. Baldini, and S. Falkow. 1984. 'Cloning of the thermostable direct or Kanagawa phenomenon-associated hemolysin of *Vibrio parahaemolyticus*', *Infect Immun*, 45: 290-2.

Kiskowski, M. A., J. F. Hancock, and A. K. Kenworthy. 2009. 'On the use of Ripley's K-function and

its derivatives to analyze domain size', *Biophys J*, 97: 1095-103.

Koch, Gary G. 1982. 'Intraclass correlation coefficient', *Encyclopedia of statistical sciences*.

Letchumanan, V., K. G. Chan, and L. H. Lee. 2014. '*Vibrio parahaemolyticus*: a review on the pathogenesis, prevalence, and advance molecular identification techniques', *Front Microbiol*, 5: 705.

Liu, X., Y. Chen, X. Wang, and R. Ji. 2004. 'Foodborne disease outbreaks in China from 1992 to 2001 national foodborne disease surveillance system (in Chinese Language)', *Journal of Hygiene Research*, 33: 725-7.

Lozano-Leon, A., J. Torres, C. R. Osorio, and J. Martinez-Urtaza. 2003. 'Identification of tdh-positive *Vibrio parahaemolyticus* from an outbreak associated with raw oyster consumption in Spain', *FEMS Microbiol Lett*, 226: 281-4.

McCarthy, SA, A DePaola, DW Cook, CA Kaysner, and WE Hill. 1999. 'Evaluation of alkaline phosphatase-and digoxigenin-labelled probes for detection of the thermolabile hemolysin (*tlh*) gene of *Vibrio parahaemolyticus*', *Letters in applied microbiology*, 28: 66-70.

McLaughlin, J. B., A. DePaola, C. A. Bopp, K. A. Martinek, N. P. Napolilli, C. G. Allison, S. L. Murray, E. C. Thompson, M. M. Bird, and J. P. Middaugh. 2005. 'Outbreak of *Vibrio parahaemolyticus* gastroenteritis associated with Alaskan oysters', *N Engl J Med*, 353: 1463-70.

McLeod, A. I. 2005. "Kendall rank correlation and Mann-Kendall trend test." In *R Package Kendall*.

Meng, W., Q. Zhong, X. Yun, X. Zhu, T. Huang, H. Shen, Y. Chen, H. Chen, F. Zhou, J. Liu, X. Wang, E. Y. Zeng, and S. Tao. 2017. 'Improvement of a Global High-Resolution Ammonia Emission Inventory for Combustion and Industrial Sources with New Data from the Residential and Transportation Sectors', *Environ Sci Technol*, 51: 2821-29.

Miyamoto, Y., T. Kato, Y. Obara, S. Akiyama, K. Takizawa, and S. Yamai. 1969. 'In vitro hemolytic characteristic of *Vibrio parahaemolyticus*: its close correlation with human pathogenicity', *J Bacteriol*, 100: 1147-9.

Molenda, J. R., W. G. Johnson, M. Fishbein, B. Wentz, I. J. Mehlman, and T. A. Dadisman, Jr. 1972. '*Vibrio parahaemolyticus* gastroenteritis in Maryland: laboratory aspects', *Appl Microbiol*, 24: 444-8.

Molero, X, RM Bartolome, T Vinuesa, L Guarner, A Accarino, F Casellas, and R Garcia. 1989. 'Acute gastroenteritis due to *Vibrio parahaemolyticus* in Spain. Presentation of 8 cases', *Medicina clinica*, 92: 1-4.

Newton, A., M. Kendall, D. J. Vugia, O. L. Henao, and B. E. Mahon. 2012. 'Increasing rates of *vibriosis* in the United States, 1996-2010: review of surveillance data from 2 systems', *Clin Infect Dis*, 54 Suppl 5: S391-5.

Nishibuchi, M., W. E. Hill, G. Zon, W. L. Payne, and J. B. Kaper. 1986. 'Synthetic oligodeoxyribonucleotide probes to detect Kanagawa phenomenon-positive *Vibrio parahaemolyticus*', *J Clin Microbiol*, 23: 1091-5.

Nishibuchi, M., and J. B. Kaper. 1995. 'Thermostable direct hemolysin gene of *Vibrio*

parahaemolyticus: a virulence gene acquired by a marine bacterium', *Infect Immun*, 63: 2093-9.

Norström, Madelaine, Dirk U Pfeiffer, and Jorun Jarp. 2000. 'A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds', *Preventive veterinary medicine*, 47: 107-19.

Paranjpye, H, L Johnson, J. M. Jacobs, and A DePaola. 2015. 'Advancing tools for modeling, forecasting and managing for *vibrio* spp. in Washington State'. https://repository.library.noaa.gov/view/noaa/16084.

R. 2000. "R Language Definition." In. Vienna, Austria: R Foundation for Statistical Computing.

Raimondi, F., J. P. Kao, C. Fiorentini, A. Fabbri, G. Donelli, N. Gasparini, A. Rubino, and A. Fasano. 2000. 'Enterotoxicity and cytotoxicity of *Vibrio parahaemolyticus* thermostable direct hemolysin in *in vitro* systems', *Infect Immun*, 68: 3180-5.

Raimondi, F., J. P. Kao, J. B. Kaper, S. Guandalini, and A. Fasano. 1995. 'Calcium-dependent intestinal chloride secretion by *Vibrio parahaemolyticus* thermostable direct hemolysin in a rabbit model', *Gastroenterology*, 109: 381-6.

Ribeiro, Paulo J., and Peter J. Diggle. 2016. "geoR: Analysis of Geostatistical Data." In *R package Version 1.7-5.2*.

Robert-Pillot, A., A. Guenole, J. Lesne, R. Delesmont, J. M. Fournier, and M. L. Quilici. 2004. 'Occurrence of the tdh and trh genes in *Vibrio parahaemolyticus* isolates from waters and raw shellfish collected in two French coastal areas and from seafood imported into France', *International journal of food microbiology*, 91: 319-25.

Songsaeng, S., P. Sophanodora, J. Kaewsrithong, and T. Ohshima. 2010. 'Quality changes in oyster (Crassostrea belcheri) during frozen storage as affected by freezing and antioxidant', *Food Chemistry*, 123: 286-90.

Stata. 2015. "Stata Statistical Software: Release 14." In. College Station, TX: StataCorp LLC.

Su, Y. C., and C. Liu. 2007. '*Vibrio parahaemolyticus*: a concern of seafood safety', *Food Microbiol*, 24: 549-58.

Sun, Z., J. Liu, S. Zhuo, Y. Chen, Y. Zhang, H. Shen, X. Yun, G. Shen, W. Liu, E. Y. Zeng, and S. Tao. 2017. 'Occurrence and geographic distribution of polycyclic aromatic hydrocarbons in agricultural soils in eastern China', *Environ Sci Pollut Res Int*, 24: 12168-75.

Van Maanen, HRE, H Nobach, and LH van Benedict. 1999. 'Improved estimator for the slotted autocorrelation function of randomly sampled LDA data', *Measurement Science and Technology*, 10: L4.

Waller, Lance A, and Carol A Gotway. 2004. *Applied spatial statistics for public health data* (John Wiley & Sons).

CURRICULUM VITAE

ZHE SUN

Fulbright Scholar, Novelist

Born on October 27, 1993 in Yuyao, Zhejiang Province, China

csuen@jhu.edu

EDUCATION

Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD.

M.Sc. in *Population Health* (2016.09-2017.12)

College of Urban and Environmental Science, Peking University, Beijing, China

B.Sc. in *Environmental Sciences* (2012.09-2016.06)

School of Mathematical Sciences, Peking University, Beijing, China

B.Sc. in *Applied Mathematics* (2013.09-2016.06)

ACADEMIC PUBLICATIONS

Zhe Sun et al. (2017*a*). Occurrence and geographic distribution of polycyclic aromatic

hydrocarbons in agricultural soils in eastern China. *Environmental Science and Pollution Research*, 24.13: 12168-12175.

Zhe Sun et al. (2017*b*). Occurrence of *nitro*- and *oxy*-PAHs in agricultural soils in eastern China and excess lifetime cancer risks from human exposure through soil ingestion. *Environment International*, 108: 261-270.